# Entrez Gene Help: Integrated Access to Genes of Genomes in the Reference Sequence Collection

Created: September 13, 2006
Updated: November 13, 2006

## Introduction

With the sequencing and annotation of key genomes, having a gene-based view of the resultant information is useful. Entrez Gene has therefore been implemented to supply key connections in the nexus of map, sequence, expression, structure, function, citation, and homology data. Unique identifiers are assigned to genes with defining sequences, genes with known map positions, and genes inferred from phenotypic information. These gene identifiers are used throughout NCBI's databases and tracked through updates of annotation and related information. Entrez Gene covers genomes represented by NCBI Reference Sequences [http://www.ncbi.nlm.nih.gov/RefSeq] (or RefSeqs) and is integrated for indexing and query and retrieval from NCBI's Entrez and E-Utilities systems.

## How Data Are Maintained

### New Records

Records are added to Entrez Gene if any of the following conditions is met:

- A RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq] is created for a completely sequenced genome and that record contains annotated genes. In the case of RNA viruses with polyprotein precursors, annotated proteins may be treated as equivalent to a ”gene”.
- A recognized genome-specific database provides information about genes (preferably with defining sequence) or mapped phenotypes.
- The NCBI Genome Annotation Pipeline [http://www.ncbi.nlm.nih.gov/genome/guide/build.html] reports model genes.
- A model organism is scheduled for sequencing, and representative sequences are identified to characterize known genes.

The minimum set of data necessary for a gene record, therefore, is: a unique identifier, or GeneID, assigned by NCBI; a preferred symbol; and either defining sequence information, map information, or official nomenclature from an authority list.

Gene records are not created for genomes that are incompletely represented by whole genome shotgun (WGS) assemblies, which are provided in terms of RefSeqs by accession numbers of the pattern NZ_ABCD12345678 [http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions]. Although not all existing records have been removed, loci defined by repetitive elements, endogenous retroviruses not named by nomenclature authorities, and loci identified by single transcripts with no other supporting data also are not in scope for Entrez Gene.

## Updates

Records are updated when new information is received. For some genomes, this may occur when a genome is re-annotated and the corresponding RefSeqs are updated. For other genomes, this may occur when any information attached to a single gene record is altered. Updates are processed daily.

Some components of the Entrez Gene record are updated automatically from other resources. Table 1 summarizes these data elements, their sources, and the update frequency. For example, GeneRIFs [http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html] are processed independently of the Entrez Gene record. Most GeneRIFs are provided by the staff of the National Library of Medicine's Index Section and are integrated weekly. Those are available with the first update to Gene of the week.

When any change is made to a record, the modification date is changed. This includes changes in GeneRIFs. The modification date, therefore, is the later of any update to Gene or supplemental information.

## Suppressed Records

Entrez Gene will suppress a record for several reasons:

- Review by NCBI staff and/or collaborators indicates that a record is no longer supported or in scope for Entrez Gene. An explanation for the suppression is provided by RefSeq staff.
- Review by NCBI staff and/or collaborators indicates that the original record defined only part of what is now understood to be the functional gene unit. In that event, one record is made secondary to another, and the URL to the current record is provided.
- The molecular basis for a Gene record that was previously only a mapped phenotype is discovered, and there was already a record for the causative locus or loci. The record for the mapped phenotype is made secondary to one of the causative loci and added to the phenotype section of all.

By default, all records, i.e., current and suppressed, are retrieved by an unqualified query. There are several methods that can be used to restrict your results to current records. These include qualifying your query with the phrase “AND alive[property]“ or using the check box Current Records in the Include Only section of the Limits page. Query Tips provides additional details.

**Table 1. Data sources for Entrez Gene.**

| Data category | Source | Species | Update frequency |
|---|---|---|---|
| Official nomenclature | HUGO Gene Nomenclature Committee (HGNC) | Human | Daily |
| | Mouse Gene Nomenclature Committee (MGNC) | Mouse | Daily |
| | Rat Gene Nomenclature Committee | Rat | Bimonthly |

| Data category | Source | Species | Update frequency |
|---|---|---|---|
| | Zebrafish Nomenclature Commitee (ZNC) | Zebrafish | Weekly |
| | FlyBase | D. melanogaster | Data release |
| GeneRIF | Index Section, NLM/public | All | Weekly/Daily |
| GO terms | Gene Ontology | Several | Weekly |
| KEGG pathway | KEGG | Several | Weekly |
| REACTOME | REACTOME | Several | Data release |

# How Data Are Displayed (Display Options)

NCBI's Entrez system supports multiple display options for each of its databases. The display options can be browsed by expanding the drop-down menu of the Display showing Summary (just above the tabs) (Figure 1B). The options represent functions that can be divided into several subcategories, which include:

- short reports of gene-specific data (Summary, Brief, Gene Table, UI List)
- comprehensive reports of gene-specific data (Full Report, ASN.1, Display XML)
- links to records in other Entrez databases based on a set of records retrieved from Entrez Gene
- links to databases outside the Entrez system, usually databases not in NCBI (LinkOut)

## Short Reports

Short reports provide details about the subsets of gene-specific data. The display options are:

- Summary
- Brief
- Gene Table
- UI List

### Summary

When you process a query, the results are displayed in the Summary format (or “docsum”) as shown in Figure 1. You can see that this is the Summary format by noting the word Summary in the Display box.

In the Summary format, each result is numbered, and a check box is provided at the left of the number. The check box enables you to select which of the records in the retrieval set that you want to review in another format, according to your selection in the Display menu. If none is checked, all are displayed in the selected format.

The text of the summary includes the preferred gene symbol (a short form of the gene name (the Official Symbol when available)), the complete name if available, the binomial name (genus species) (in brackets), other symbols and names, other designations, the genomic location (Chromosome, Location), the Mendelian Inheritance in Man (MIM) number for the gene (human only), and the GeneID (Figure 1C). If the gene is on a named plasmid, then the plasmid name is given as the location. The Links menu, at the right of all displays, provides related records in Entrez. The calculation of these links is documented in the web page accessible from the link anchored by a question mark (?) at the upper left corner of the expanded menu (Figure 2A). If you navigate to that documentation on the web, click on Gene to navigate quickly to the description of gene-specific links. The navigation in the Links menu (attached to individual records) is based on the same infrastructure in Entrez that supports navigation to records related to a set of query results. More complete documentation of each type of link in this document is in the section entitled Finding Data Related to Entrez Gene in Other Entrez Databases.

## Brief

The functions allowed from the Brief display are similar to those described from the Summary display. The purpose of the Brief option is to support a more compact result set while providing enough information (the preferred symbol, 20 characters of the full name, and the GeneID) for you to select records to display a Full Report.

## Gene Table

The Gene Table display represents the gene structure as annotated on the current genomic RefSeq representing the reference genome. This annotation is updated only with each comprehensive re-annotation cycle at NCBI (a build [http://www.ncbi.nlm.nih.gov/genome/guide/build.html]), usually no more than twice a year. This means that if the version of a RefSeq RNA has changed, the table will be out of date.

The table reports the intron/exon organization of each transcript, and, if an mRNA, the region of each exon that contains coding sequence (sample). It does this in two ways:

- graphically, by repeating the display included in the Full Report
- in a table, by reporting the position of any exon, intron, or coding region

In this display, you can browse the structure of any gene and its products, and this view makes it easier download the gene-related sequence, as summarized in Table 2.

Please note that this function is not supported when the gene has not (yet) been annotated on any of NCBI's Genomic RefSeqs.

The sequence being retrieved is from the indicated genomic sequence, not the RNA. This means that the length of a poly(A) tail is not included in the report.

When viewing the sequence-specific nucleotide or protein record, use the Display options to generate the format you prefer.

Because the Gene Table reflects the annotation on the current genomic sequence, for bulk access you may prefer to use the seq_gene.md.gz file in the species_specific mapview subdirectory. These files are available for genomes that can be viewed in Map Viewer. For example:

- Human [ftp.ncbi.nih.gov/genomes/H_sapiens/mapview]
- Mouse [ftp.ncbi.nlm.nih.gov/genomes/M_musculus/mapview]

The data in Gene Table reflect the current NCBI batch annotation, so it is possible that the RefSeq mRNAs, which are updated continuously, have changed. Check also the Reference Sequences section of the Entrez Gene record to determine whether updates have occurred (new versions and/or more variants and/or temporary suppression during a review process).

**Navigation**: In the Genomic regions, transcripts, and products section, the mRNA and Protein links in the table are hyperlinked to the Exon information table specific to that accession.

To the right of the Exon information tables are blue-shaded icons:

- ▲ top of the page
- ◀ table for the previous product
- ▶ table for the next product

Please see Table 2 for information on access to sequence information from the Gene Table display option.

## UI List

This display is essentially the same as that of the Brief format, with the addition of the unique identifiers (UIs) for a Gene record on the second line. The difference between the Brief and UI List displays is more apparent, however, when selecting the Send to Text option. For the UI List, only the Gene identifiers are reported.

# Full Reports

All of the content that Entrez Gene provides is defined by the ASN.1 file. The Full Report display is of the HTML transformation of that ASN.1 and includes navigation tools (Table of Contents and Links), diagrams, and text. Some gene-specific information is **not** maintained in Entrez Gene but is maintained in more specialized databases such as GEO [http://www.ncbi.nlm.nih.gov/projects/geo], HomoloGene, UniGene, and Probe. Access to the additional information maintained in other resources within NCBI or external to NCBI is provided by the Links menu at the top right and by other HTML anchors within the page.

The Full Report display is divided into the gray Search bar with tabs (explained in Query tips, Figure 1B), the gray Display bar with tabs (Figure 1C), and then according to the major subsections of an Entrez Gene report.

- Navigation Menu
- Title
- Summary
- Genomic Regions, Transcripts, and Products
- Genomic Context
- Bibliography

- Interactions
- Alleles
- General Gene Information
- General Protein Information
- NCBI Reference Sequences (RefSeqs)
- Related Sequences
- Additional Links

For convenience, there is another Display bar at the bottom of the page. Some options discussed here are:

- ASN.1
- XML

If multiple records are selected for display, the start of each record is indicated by the numbered open box at the left and by the Links menu at the right.

## Navigation Menu

The menu at the right of the Entrez Gene report supports navigation to multiple sites of interest (Figure 2). In some display formats, the menu can be expanded and compressed by clicking on the down (▼) or right (▶) arrows, respectively. More details about each submenu follow.

Table of Contents lists the subcategories of information available for a gene. Clicking on the name of the subcategory takes you to that portion of the gene record. The arrow pointing up on the bar separating subcategories (⬆) will return you to the top of the page, should you want to make a different selection from the menu.

Links contains standard links seen in all Entrez records, followed by links to sites of interest not in the Entrez system. The LinkOut option is always last in this section when available.

Entrez Gene Info enumerates resources that may help you find and understand the information in Gene. The Help link goes to the default help document. The default help document is also accessed by the question marks (?) in the horizontal section separators.

Feedback enumerates several sites where you can comment on or add data to Gene and/or RefSeq.

Subscriptions provides links to forms where you can subscribe to a mailing list to receive announcements about updates to Gene, Map Viewer, and RefSeq.

## Title

The first line below the second row of tabs is referred to as the “Title” (Figure 3A) and provides the preferred symbol and descriptive name in bold font, followed by the italicized binomial in brackets. If there is a recognized authority for the gene nomenclature of a species, then that authority is the source for these values.

The second line of the title section contains the NCBI GeneID and the last date a record was changed. The date is in the format day-month-year. Change is defined as any modification to the content of the record, including ancillary changes such as the URL for a displayed link.

## Summary

The Summary section of the Full Report display (Figure 4) includes several categories of information for each gene as available. These include:

Official Symbol: and Name: Nomenclature provided by the named external authority.

Identifier from the primary data provider: Identifier and link to the major resource outside of NCBI that provided information about this gene.

Locus tag for the record is in the next line. Locus tag corresponds to the systematic feature [http://www.ncbi.nlm.nih.gov/projects/collab/FT/index.html#qual_summary] qualifier used by the international sequence collaboration (DDBJ/EMBL/GenBank) and can be assigned by sequence submitters as a unique, systematic gene descriptor. When such a value is not available from submitted sequence, the identifier from a collaborating model organism database is used. Locus tag is often used to anchor a link to a database other than Entrez Gene.

Gene type: Possible values are tRNA, rRNA, snRNA, scRNA, snoRNA, miscRNA, protein-coding, pseudo, other, and unknown. These are indexed as properties of a gene by using the terms:

- genetype trna (transfer RNA, tRNA)
- genetype rrna (ribosomal RNA, rRNA)
- genetype snrna (small nuclear RNA, snRNA)
- genetype scrna (small cytoplasmic RNA)
- genetype snorna (small nucleolar RNA)
- genetype miscrna (miscellaneous RNA)
- genetype protein coding
- genetype pseudo (pseudogene)
- genetype other (when the type is known, but there is no specific enumeration for it)
- genetype unknown (when the type of gene is uncertain)

RefSeq status: Any of the set of status descriptions defined by RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status].

Organism: The binomial, and strain when appropriate, with a link to the NCBI Taxonomy database.

Lineage: Binomial and lineage from the Taxonomy database.

Also known as: Unofficial symbols and descriptions that have been used for this gene and its products. If there is no official symbol, the symbol at the top of the display is repeated in this section. These names are integrated from several sources, including model organism databases, annotation on sequence records, and interactive curation from the published literature.

Summary: Descriptive text about the gene, its cellular localization, its function, and its effect on phenotype.

## Genomic Regions, Transcripts, and Products

This portion of the Full Report (Figure 4B) is provided when a gene has been annotated on a genomic RefSeq, in other words, when the position of a pseudogene, or the intron/exon/coding region information, is available in some genomic coordinate system. You can use this section to:

- view the intron/exon/coding region organization of a gene and its RNA product, or the placement of a pseudogene, on a genomic RefSeq
- identify the RefSeqs that correspond to any RNA or protein product and see an overview of the exons they represent
- navigate to the genomic, RNA, or protein sequence, by clicking on the RefSeq accession
- navigate to Blink, from the protein accession, to review how this protein sequence compares to related proteins from other taxa

Each position of a gene product, when represented by a RefSeq RNA (accession NM_000000? NM_000000000 or XM_000000/XM_000000000 for mRNA, NR_000000 or XR_000000 for non-protein coding RNAs) and/or protein (NP_000000/NP_000000000 or XP_000000/XP_000000000), is provided relative to the genomic accession on which it is annotated. Each accession is an anchor to a menu that will display the sequence in several formats. Protein accession numbers also facilitate retrieval of specific BLink, CDD, or COG displays. See the following figure as an example, and read the detailed description in Box 1. Interactive View. Be sure to click the link in the box for the interactive view.

For some species, including human and other vertebrates, the genomic RefSeqs are updated independently of the annotated product RNAs, with the latter being updated more frequently. This means that several kinds of discrepancies between the diagram and the current RefSeq RNAs may result.

- The diagram may be labeled with an mRNA accession (for a predicted transcript) of the format XM_123456, yet clicking on that accession results in an entry in Entrez Nucleotide that indicates that this accession is no longer primary. That means that a curated mRNA (accession of the format NM_123456 or NM_123456789) has been generated to replace the previous model accession. This new "NM" accession will be reported in the Reference Sequences section.
- The diagram may be labeled with curated RNA accession numbers (of the format NM_123456 or NM_123456789 or NR_123456) different from those listed in the RefSeq section. This will result if curation after the submission of the annotated genome identified more transcript variants, which therefore are listed only in the Reference Sequence section but not in the diagram. It will also result if curation after submission of the annotated genome identified an error in the annotated product, and the accession for that product was suppressed. In that case, the Genomic regions, transcripts and products section will indicate a transcript not listed in the RefSeq section of the Entrez Gene report.

## Genomic Context

The Genomic context (Figure 5A) section reports the location of the gene on the chromosome in non-sequence coordinates and the strain and genotype information of the source sequence. The title bar of this section includes a link to Map Viewer, providing the same display as that generated from the Map Viewer link in the Links menu.

If the gene has been included in a genomic annotation, the section also diagrams neighboring genes and indicates their orientations. If the name of a gene is too long to use for a label, truncation is indicated by an ellipsis (...). The gene being shown on the diagram is in maroon. All other diagrams and labels anchor links to specific gene pages, supporting quick navigation to review neighboring genes by clicking in the area of the symbol/arrow.

If a gene has not been included in the current version of the annotated genome provided in NCBI RefSeqs, the Genomic context section will not include a diagram but will report the map location. If the gene is annotated on more than one genomic RefSeq, only one will be used for the graphic display, but the location information will be provided in the ASN.1 of the record and in the Reference Sequences Section.

## Bibliography

The Bibliography section (Figure 5B) provides a link to PubMed in the same format as that generated from the PubMed option in the pull-down Links menu at the upper right portion of the page.

If GeneRIFs [http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html] (Figure 5C) have been submitted, they are included in this section. The majority of these annotations have been provided by a collaboration between the NLM's Index Section and NCBI. The GeneRIF homepage [http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html] provides more information about the project, including how general users can make submissions. Because there can be a large number of GeneRIFs, they are provided in a scrolling window.

## Interactions

There are two major subcategories of information reported as Interactions: HIV-1 interactions and general interactions.

### *HIV-1 Interactions*

The HIV-1, Human Protein Interaction Database [http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions] is funded by the Division of Acquired Immunodeficiency Syndrome (DAIDS) [http://www.niaid.nih.gov/daids] of the National Institute of Allergy and Infectious Diseases (NIAID) [http://www3.niaid.nih.gov]. As the title indicates, this project focuses on the human proteins that have been shown to interact with proteins from HIV-1. The format of this section is different for the human and HIV-1 gene reports. On human, the display consists of:

- the HIV-1 protein, which anchors a link to Entrez Gene for that gene product
- a concise description of the interaction
- links to papers in PubMed that support the described interaction

For HIV-1, the display is subdivided by peptide name and includes:

- a key word categorizing the interaction
- the full name of the human gene, which anchors a link to that record
- links to papers in PubMed that support the described interaction

Please note that there are separate reports from this section that are available for download, both from the HIV Interactions homepage and the GeneRIF subdirectory of the Gene FTP site [ftp.ncbi.nlm.nih.gov/gene/GeneRIF].

### General Interactions

Interactions in this general section are reported as pairs. The report will always include, in the first column, the product of the gene that is part of the interaction in the first column. Depending on the type of interaction, the rest of the display may report:

- the other interactant, anchoring a link to more information
- the gene name of the other interactant, anchoring a link to that record in Entrez Gene
- the complex to which the interactant(s) belongs
- the source of these data, anchoring a link to the record at that source
- a concise description of the interaction
- links to papers in PubMed that support the described interaction

## Alleles

This section reports the general characteristics of alleles that have been described for a gene and provides links to more detailed information. This function is being phased in gradually; the current set is for mouse and is being developed from information supplied by Mouse Genome Informatics [http://www.informatics.jax.org].

## General Gene Information

This section includes several subcategories of information, including:

GeneOntology (GO): The specific GO terms are listed by source of the information, category, term, evidence information, and links to supporting publications. Each GO term supports a link to the AmiGO [http://godatabase.org/cgi-bin/go.cgi] browser. Abbreviations in the Evidence column indicate the level of support for assigning a GO term to a gene. Explanations for these abbreviations are provided by the Gene Ontoloogy website [http://www.geneontology.org/GO.evidence.codes.shtml].

Entrez Gene does not alter the associations provided by a model organism database, nor does Entrez Gene recapitulate the directed acyclic graph structure provided by GO. Thus, Entrez Gene does not support retrieval of all genes associated with a specific GO term based on that term's parent.

Homology: A partial listing, with links, of orthologs in other species. Other views of homology data are available from TaxPlot [http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi?] and the HomoloGene link in the Links menu.

Phenotypes: A description of the effect of the gene on phenotype, especially disease. Links to more information are provided as available, as in the case of human disease and links to OMIM.

Markers: An enumeration of the markers that are related to this gene (Figure 6). The relationship is reported based either on direct reports, e-PCR using mRNA templates, or e-PCR-based localization on the genome within a region beginning 2 kb upstream of the gene and ending 0.5 kb downstream. Links are provided in the NCBI UniSTS database.

Pathways: A description of pathways that include this gene with links to more information about that pathway.

Relationships: At present, used for gene models to describe some of the related sequences used to support the model transcript.

## General Protein Information

This section applies only to genes that encode proteins. It reports the name or names that have been assigned to proteins encoded by the gene and provides other descriptive text. The names are as annotated on the RefSeq protein, when that protein is available. The sources of these names include model organism databases, annotation on public sequence databases, and curation by RefSeq staff.

## NCBI Reference Sequences (RefSeqs)

This section describes the gene-specific NCBI reference sequences (RefSeqs [http://www.ncbi.nlm.nih.gov/RefSeq]) that have been established for this gene. In addition to enumerating the accession numbers and providing links to the appropriate Entrez sequence database, this section may also include descriptions of each transcript variant, accession numbers of the public sequences used to support any transcript, and a listing of computed domains in an encoded protein. The text provided in this section therefore supports retrieving gene records based on descriptions of conserved domains.

The Reference Sequence group uses several approaches in maintaining information. These can be broadly categorized as:

1. *RefSeqs maintained independently of Annotated Genomes*(Figure 7). RefSeq RNA and protein sequences are updated continuously, independently of any comprehensive reannotation of a genome. Because these reference sequences are curated independently of the genome annotation cycle, their versions may not match the RefSeq versions in the current genome build. You can identify updates by comparing versions in this section to versions in the Genomic regions, transcripts, and products section.

2. *RefSeqs of Annotated Genomes* (Figure 8). This section reports genomic RefSeqs from all assemblies on which this gene is annotated, such as RefSeqs for chromosomes and scaffolds (contigs) from both reference and alternate assemblies. The position and strand of the gene feature is provided (offset 0) .GenBank and FASTA anchor links to sequence in the given formats. Model RNAs and proteins are also reported here.

3. *Genome Annotation*. RefSeq RNA and protein sequence are provided only through the process of genome/chromosome annotation.

4. *Suppressed Reference Sequence(s).* Accession numbers listed in this section were suppressed for the cited reason(s). Suppressed RefSeqs do not appear in BLAST databases, related sequence links, or BLAST links (BLink) but may still be retrieved by from the Nucleotide or Protein databases, and by clicking on the hyperlinked accession.version.

## Related Sequences

This section has two subsections, one in which the nucleotide sequence is primary and one for protein sequences only (UniProt). It contains sequence accession numbers that are related to the gene and provides links to the appropriate sequence record in Entrez Nucleotide or Entrez Protein. It is not intended to be a comprehensive list of all sequences related to any gene; such sequences can more explicitly be found by using BLAST to query sequence databases or by using pre-calculated reports of related sequences via Entrez Nucleotide, Entrez Protein, or BLink. The sequence accession numbers in this section are provided in a tab-delimited format in the gene2accession.gz file in the DATA [ftp.ncbi.nlm.nih.gov/gene/DATA] directory of the Gene FTP site.

Gene purposely lists protein accession numbers on records being represented as not protein-coding. The intent is to make the connection between sequence annotation and Gene's current representation of the type of gene. Users with evidence indicating that the Gene record should be reviewed are encouraged to contact the RefSeq staff [http://www.ncbi.nlm.nih.gov/RefSeq/update.cgi].

## Additional Links

This section provides a printable view of a subset of links to information both within and external to NCBI. Some of these links overlap those included in the Links menu. The intent of this section is to provide a printable report of, for example, MIM numbers, UniGene cluster numbers, and family-specific websites.

# Display Bar

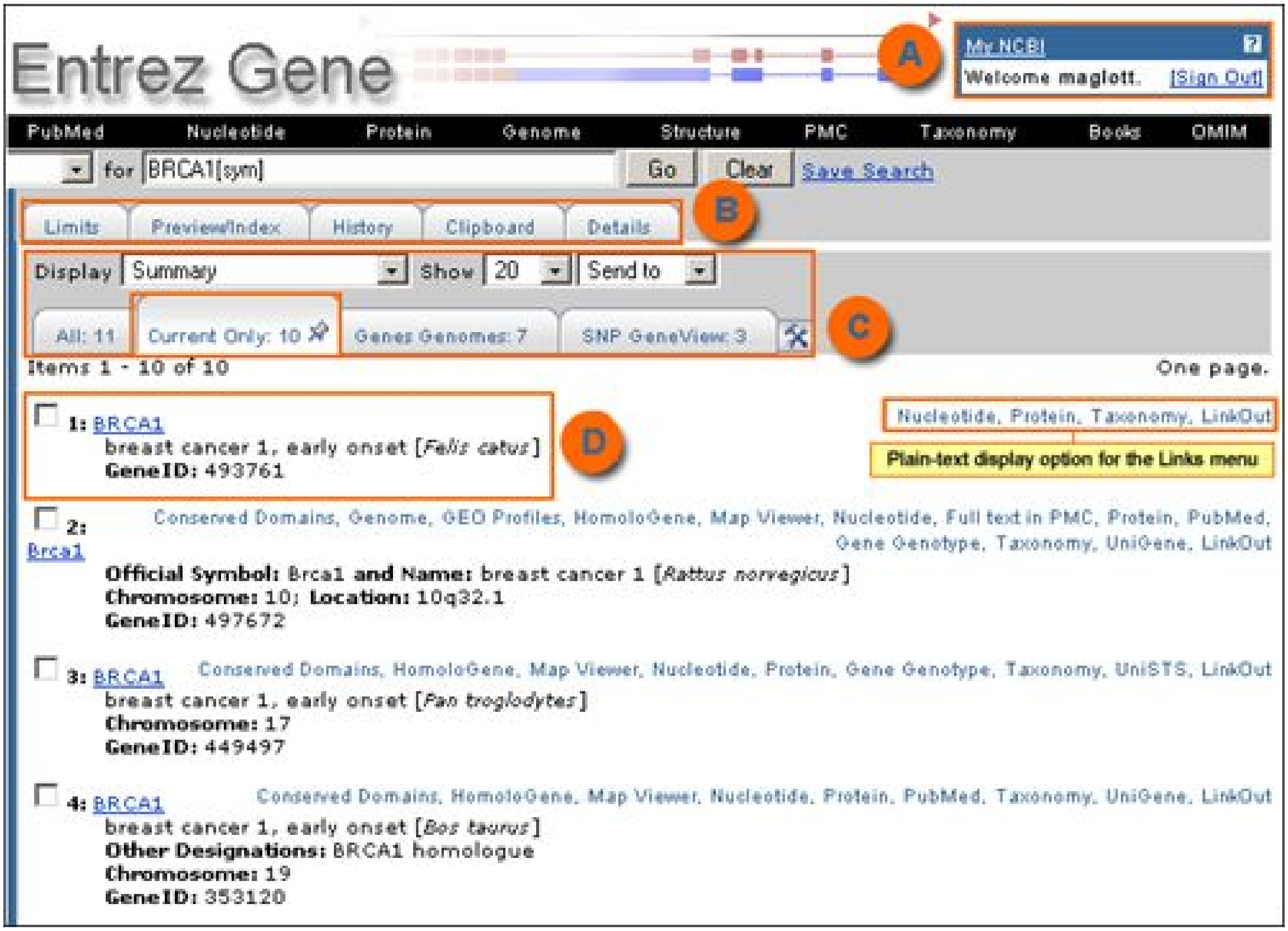
At the bottom of the Full Report is another Display bar. This bar is the same as the one at the top. There are many display options available. Of those provided, ASN.1 and XML are discussed below.

## ASN.1

The ASN.1 display provides gene records structured according to the Entrezgene specification [http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/entrezgene/entrezgene.asn]. An XML transformation of the ASN.1 is also available. Detailed information about the specification is provided in the Tips for Programmers section.

## XML

Any record or selected set of records can be displayed in XML format. The XML is generated automatically from the ASN.1 record that is used to support the display, with the names of the tags defined by the ASN.1 specification. Detailed information about the specification is provided in the Tips for Programmers section.



for BRCA1[sym]

Display Summary Show 20

□ 1: BRCA1
breast cancer 1, early onset [*Felis catus*]
**GeneID:** 493761

□ 2: Brca1 Conserved Domains, Genome, GEO Profiles, HomoloGene, Map Viewer, Nucleotide, Full text in PMC, Protein, PubMed, Gene Genotype, Taxonomy, UniGene, LinkOut
**Official Symbol:** Brca1 **and Name:** breast cancer 1 [*Rattus norvegicus*]
**Chromosome:** 10; **Location:** 10q32.1
**GeneID:** 497672

□ 3: BRCA1 Conserved Domains, HomoloGene, Map Viewer, Nucleotide, Protein, Gene Genotype, Taxonomy, UniSTS, LinkOut
breast cancer 1, early onset [*Pan troglodytes*]
**Chromosome:** 17
**GeneID:** 449497

□ 4: BRCA1 Conserved Domains, HomoloGene, Map Viewer, Nucleotide, Protein, PubMed, Taxonomy, UniGene, LinkOut
breast cancer 1, early onset [*Bos taurus*]
**Other Designations:** BRCA1 homologue
**Chromosome:** 19
**GeneID:** 353120

Figure 1. A representative Summary report from Entrez Gene. Representative Summary display resulting from a query (note the text in the box labeled for in the gray query bar) for records having **BRCA1** as the symbol. This figure is a screen captured when the user maglott was logged into My NCBI**(A)** and had selected the plain text display option from the Links menu. (A representative expandable menu from Entrez Gene is shown in Figure 2A tabs.) The use and implementation of the set of tabs (Limits, Preview/Index, History, Clipboard, Details) in the upper gray area in the Summary and all other display formats **(B)** are common to Entrez databases and are documented in detail [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp]. The second gray bar **(C)** also has multiple tabs. These serve multiple functions, primarily to (1) show the counts of some subcategories of records in the result set and (2) display only the records in the subcategory that is indicated. In this example, the tab Current Only was selected so that 10 records are now displayed in the summary. **(D)** The summary of the query results. To see one of the entries, click on its gene symbol. To see one or more entries: check one or more of the open boxes to the left of the symbols; use the menu options in the Display box to select Full Report; and press Display. The text of the each Summary display includes the gene symbol, the full descriptive name (indicated as Official if provided by an authoritative nomenclature group), the binomial for the species, the chromosome and regional localization, other names, the MIM number if appropriate, and the GeneID value.

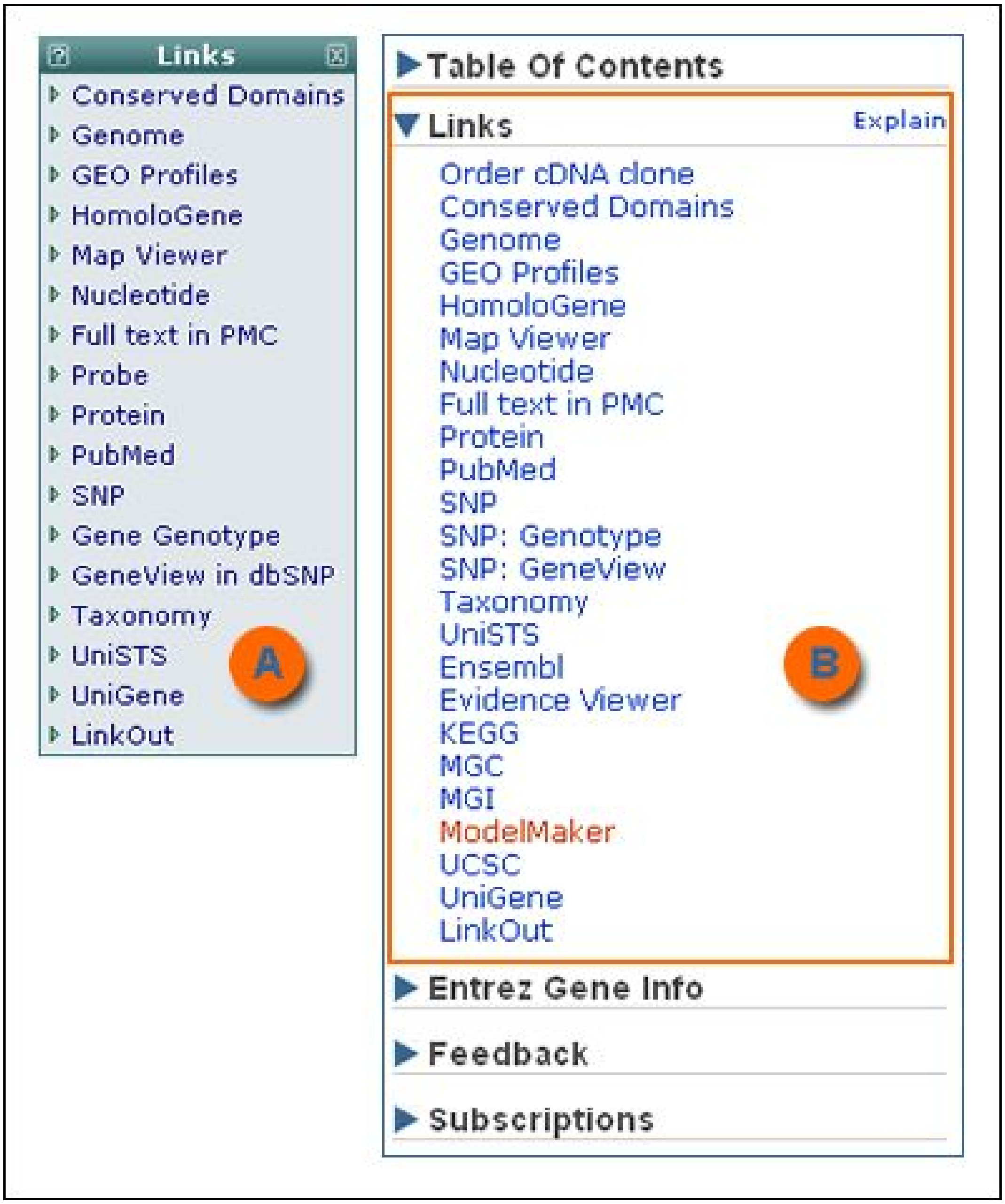
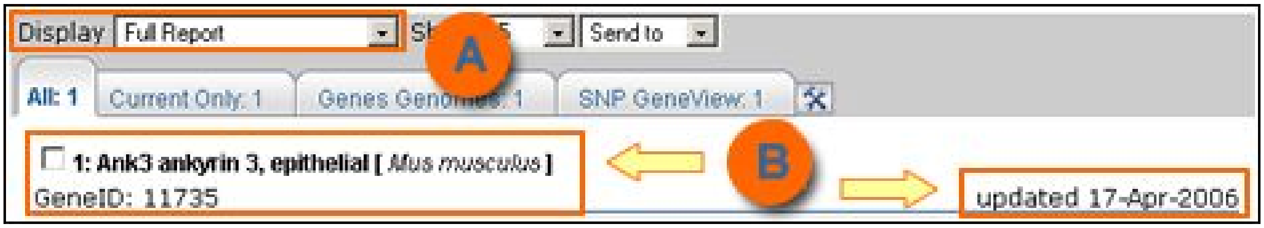


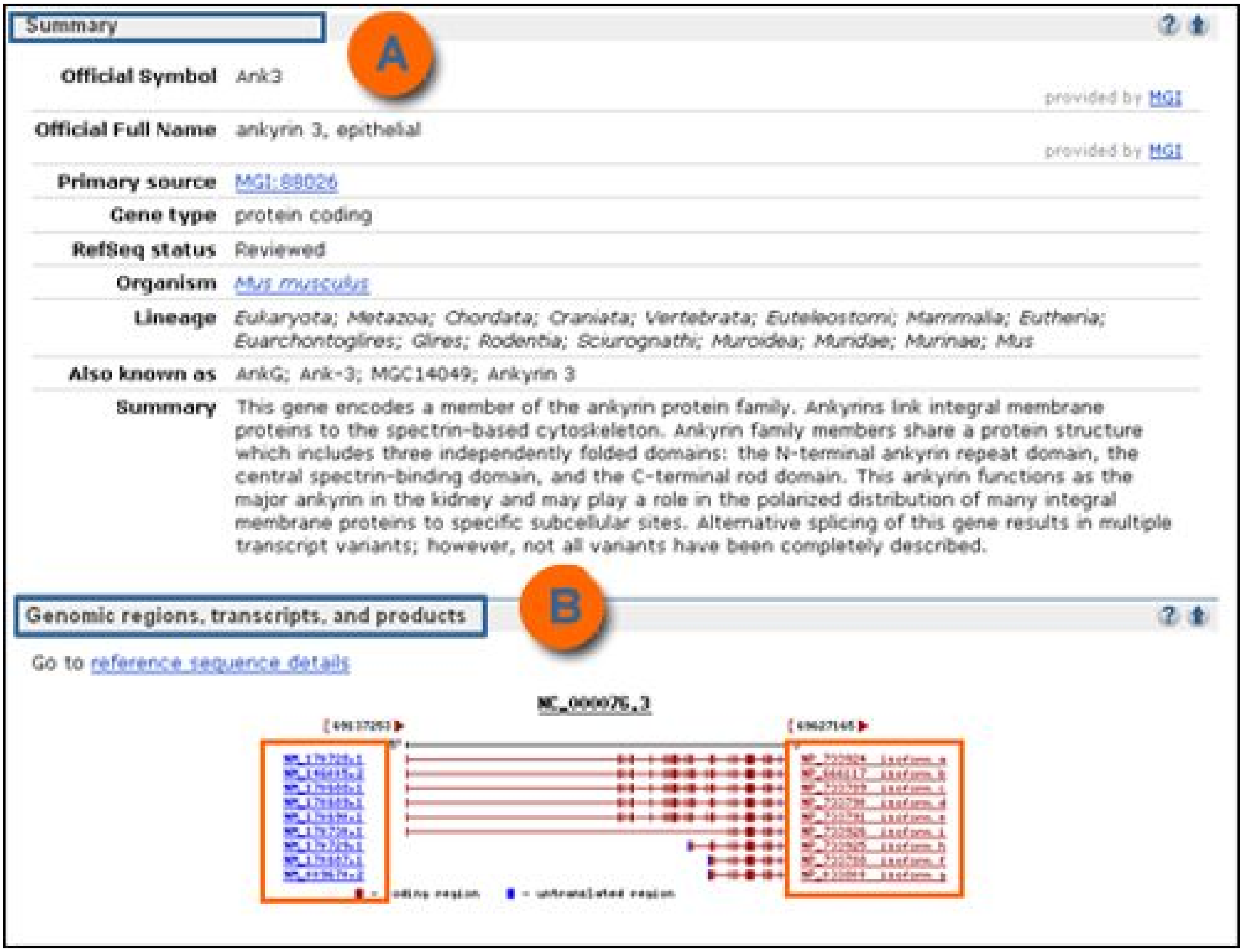
Figure 2. Representation of Links in Entrez Gene. Links to additional information are provided in multiple formats and with different content. In the summary views (Summary, Brief), the word Links at the upper right of the page expands to a navigation menu to other NCBI databases **(A)**. If you have configured your My NCBI environment to display the menu as plain text, the menu will be written out as in Figure 1. Within the Full Report and Gene Table display options, the menu also includes includes a summary of the links to non-NCBI resources mentioned in the record **(B)**.



□ 1: Ank3 ankyrin 3, epithelial [ *Mus musculus* ]
GeneID: 11735

updated 17-Apr-2006

Figure 3. Representative title section of the Full Report display **(A)**. This section provides the symbol and full name of the gene, the GeneID, and the last date any information in this record was changed **(B)**. See Figures 4–8 for subsequent section views of the Full Report display.



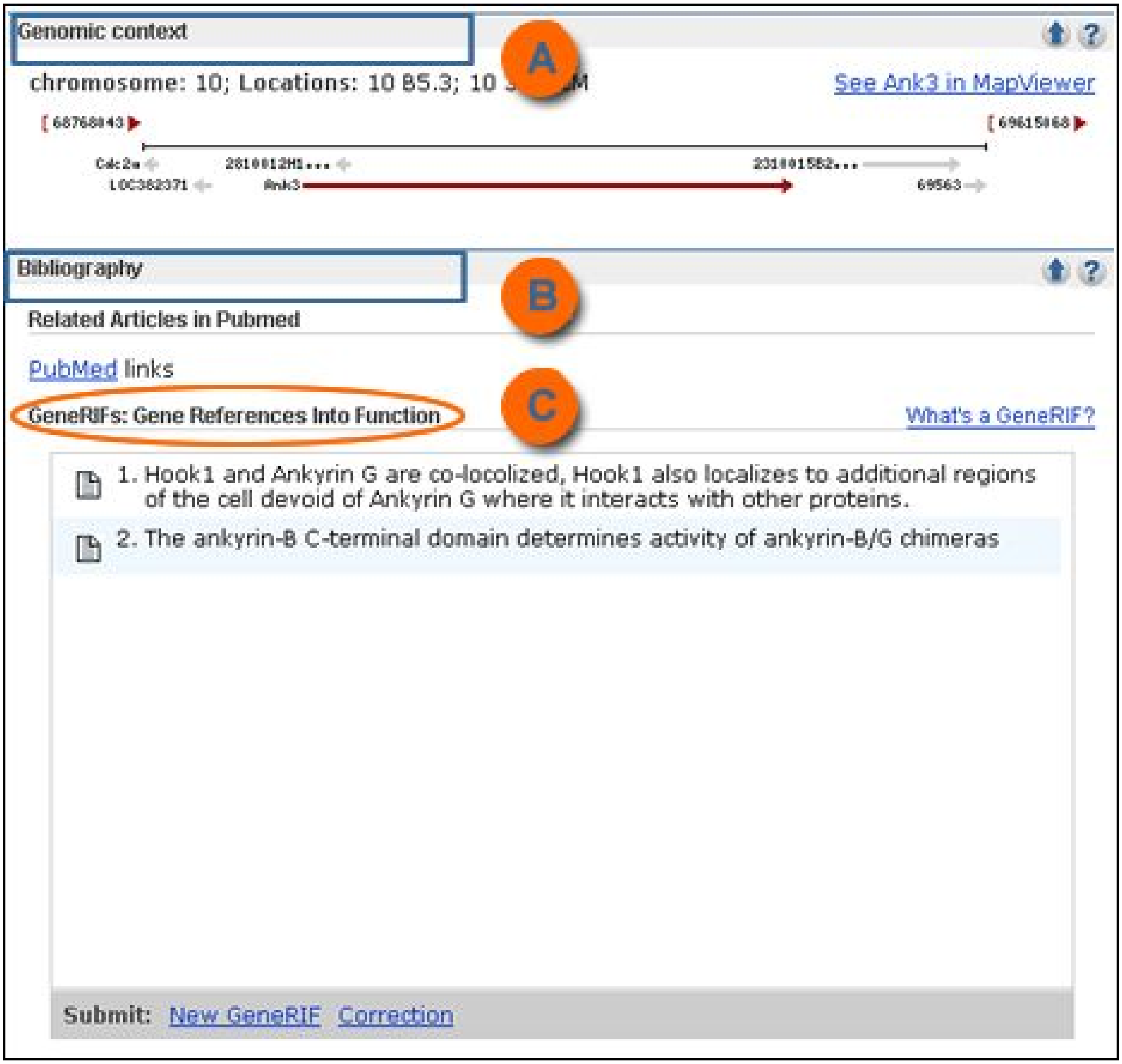
Summary

| | |
|---|---|
| **Official Symbol** | Ank3<br>provided by MGI |
| **Official Full Name** | ankyrin 3, epithelial<br>provided by MGI |
| **Primary source** | MGI:88026 |
| **Gene type** | protein coding |
| **RefSeq status** | Reviewed |
| **Organism** | *Mus musculus* |
| **Lineage** | *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus* |
| **Also known as** | AnkG; Ank-3; MGC14049; Ankyrin 3 |
| **Summary** | This gene encodes a member of the ankyrin protein family. Ankyrins link integral membrane proteins to the spectrin-based cytoskeleton. Ankyrin family members share a protein structure which includes three independently folded domains: the N-terminal ankyrin repeat domain, the central spectrin-binding domain, and the C-terminal rod domain. This ankyrin functions as the major ankyrin in the kidney and may play a role in the polarized distribution of many integral membrane proteins to specific subcellular sites. Alternative splicing of this gene results in multiple transcript variants; however, not all variants have been completely described. |

Genomic regions, transcripts, and products

Go to reference sequence details

Figure 4. Representative Summary**(A)** and Genomic regions, transcripts, and products sections **(B)** of a Full Report display. Each accession in this display is hyperlinked to the Nucleotide or Protein databases as appropriate, where standard tools exist to facilitate downloading sequence.



Genomic context

chromosome: 10; Locations: 10 B5.3; 10 [illegible] M

See Ank3 in MapViewer

Bibliography

Related Articles in Pubmed

PubMed links

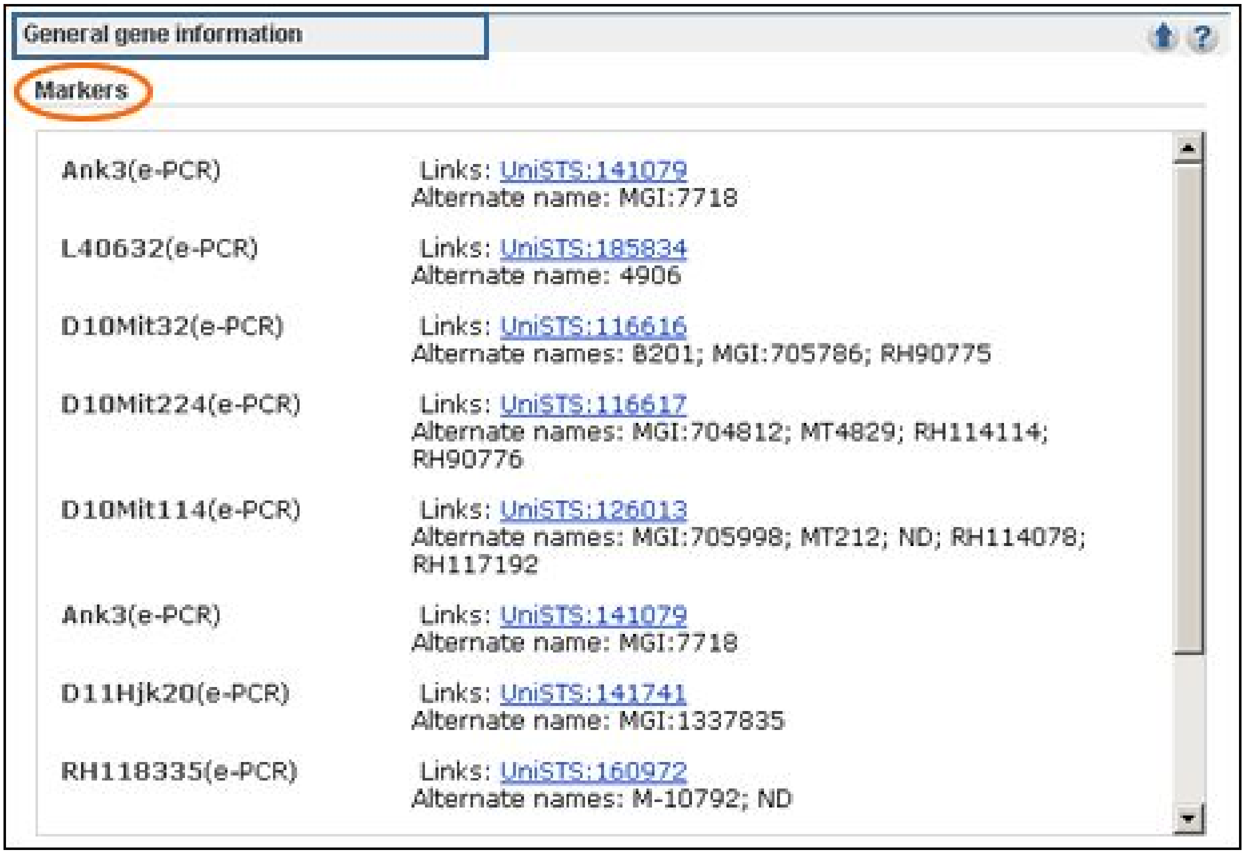
GeneRIFs: Gene References Into Function

What's a GeneRIF?

1. Hook1 and Ankyrin G are co-locolized, Hook1 also localizes to additional regions of the cell devoid of Ankyrin G where it interacts with other proteins.
2. The ankyrin-B C-terminal domain determines activity of ankyrin-B/G chimeras

Submit: New GeneRIF Correction

Figure 5. Representative Genomic context**(A)** and Bibliography**(B)** sections of a Full Report display. Note that the GeneRIFs**(C)** subsection is provided in a scrolling window.

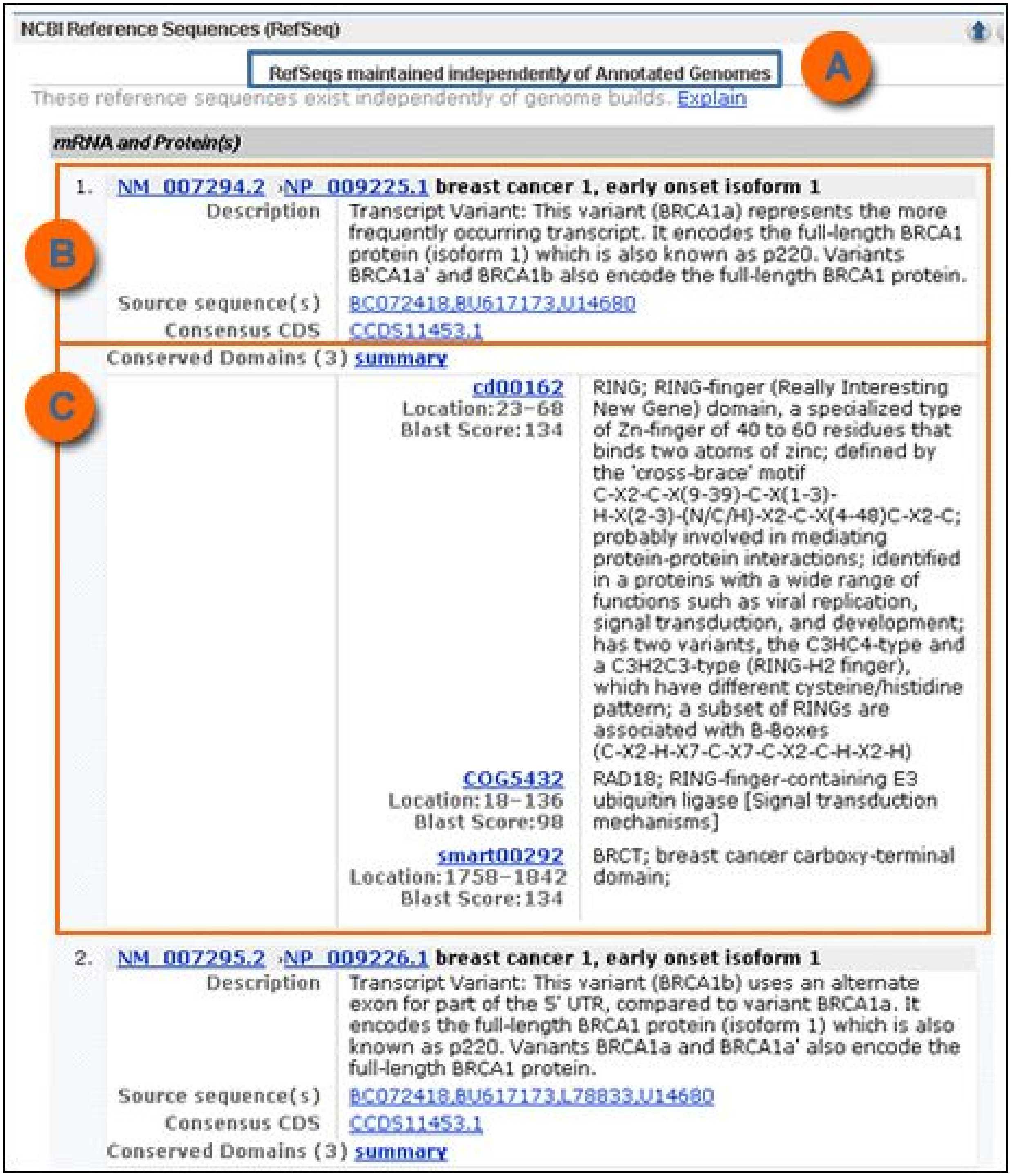


General gene information

Markers

| | |
|---|---|
| Ank3(e-PCR) | Links: UniSTS:141079<br>Alternate name: MGI:7718 |
| L40632(e-PCR) | Links: UniSTS:185834<br>Alternate name: 4906 |
| D10Mit32(e-PCR) | Links: UniSTS:116616<br>Alternate names: B201; MGI:705786; RH90775 |
| D10Mit224(e-PCR) | Links: UniSTS:116617<br>Alternate names: MGI:704812; MT4829; RH114114; RH90776 |
| D10Mit114(e-PCR) | Links: UniSTS:126013<br>Alternate names: MGI:705998; MT212; ND; RH114078; RH117192 |
| Ank3(e-PCR) | Links: UniSTS:141079<br>Alternate name: MGI:7718 |
| D11Hjk20(e-PCR) | Links: UniSTS:141741<br>Alternate name: MGI:1337835 |
| RH118335(e-PCR) | Links: UniSTS:160972<br>Alternate names: M-10792; ND |

Figure 6. Representative General gene information section of a Full Report display. Note that the Markers subsection is provided in a scrolling window.



NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

*mRNA and Protein(s)*

1. NM_007294.2 →NP_009225.1 **breast cancer 1, early onset isoform 1**

Description: Transcript Variant: This variant (BRCA1a) represents the more frequently occurring transcript. It encodes the full-length BRCA1 protein (isoform 1) which is also known as p220. Variants BRCA1a' and BRCA1b also encode the full-length BRCA1 protein.

Source sequence(s): BC072418,BU617173,U14680

Consensus CDS: CCDS11453.1

Conserved Domains (3) summary

| | |
|---|---|
| cd00162<br>Location:23–68<br>Blast Score:134 | RING; RING-finger (Really Interesting New Gene) domain, a specialized type of Zn-finger of 40 to 60 residues that binds two atoms of zinc; defined by the 'cross-brace' motif C-X2-C-X(9-39)-C-X(1-3)-H-X(2-3)-(N/C/H)-X2-C-X(4-48)C-X2-C; probably involved in mediating protein-protein interactions; identified in a proteins with a wide range of functions such as viral replication, signal transduction, and development; has two variants, the C3HC4-type and a C3H2C3-type (RING-H2 finger), which have different cysteine/histidine pattern; a subset of RINGs are associated with B-Boxes (C-X2-H-X7-C-X7-C-X2-C-H-X2-H) |
| COG5432<br>Location:18–136<br>Blast Score:98 | RAD18; RING-finger-containing E3 ubiquitin ligase [Signal transduction mechanisms] |
| smart00292<br>Location:1758–1842<br>Blast Score:134 | BRCT; breast cancer carboxy-terminal domain; |

2. NM_007295.2 →NP_009226.1 **breast cancer 1, early onset isoform 1**

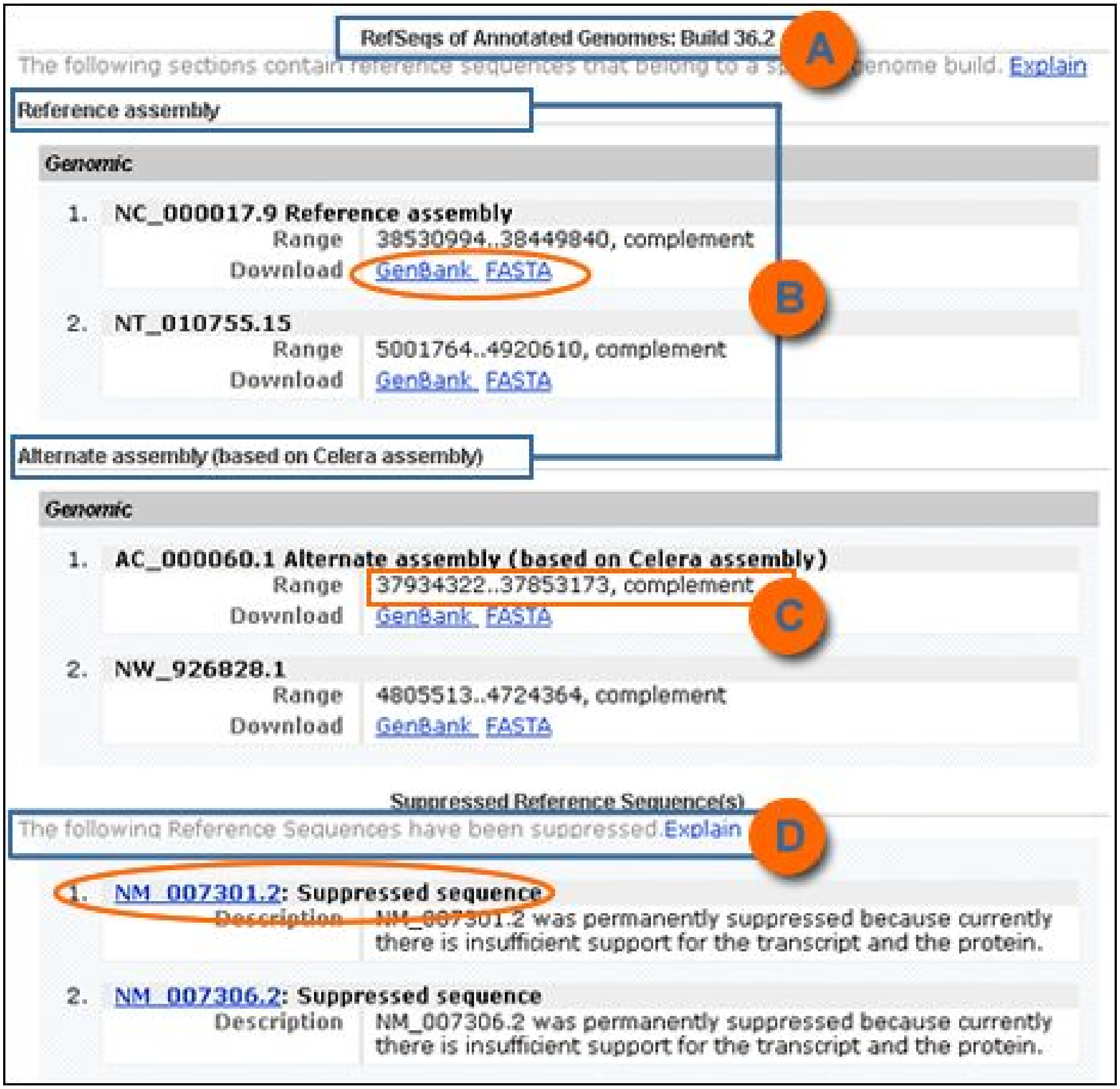
Description: Transcript Variant: This variant (BRCA1b) uses an alternate exon for part of the 5' UTR, compared to variant BRCA1a. It encodes the full-length BRCA1 protein (isoform 1) which is also known as p220. Variants BRCA1a and BRCA1a' also encode the full-length BRCA1 protein.

Source sequence(s): BC072418,BU617173,L78833,U14680

Consensus CDS: CCDS11453.1

Conserved Domains (3) summary

Figure 7. Representative NCBI Reference Sequences (RefSeq) section in the Full Report display. This section includes two subsections: RefSeqs maintained independently of Annotated Genomes**(A)**, and RefSeqs of Annotated Genomes (see Figure 8). The RefSeqs maintained independently of Annotated Genomes includes: **B,**mRNA and Protein(s) accession numbers, a Description of the transcript, links to the Source sequence(s) from which the Reference Sequences are derived, and links to the Consensus CoDing Sequence database (Consensus CDS); and **C,** links to the Conserved Domains database.



RefSeqs of Annotated Genomes: Build 36.2

The following sections contain reference sequences that belong to a sp enome build. Explain

Reference assembly

Genomic

1. NC_000017.9 Reference assembly
Range 38530994..38449840, complement
Download GenBank FASTA

2. NT_010755.15
Range 5001764..4920610, complement
Download GenBank FASTA

Alternate assembly (based on Celera assembly)

Genomic

1. AC_000060.1 Alternate assembly (based on Celera assembly)
Range 37934322..37853173, complement
Download GenBank FASTA

2. NW_926828.1
Range 4805513..4724364, complement
Download GenBank FASTA

Suppressed Reference Sequence(s)

The following Reference Sequences have been suppressed. Explain

1. NM_007301.2: Suppressed sequence
Description NM_007301.2 was permanently suppressed because currently there is insufficient support for the transcript and the protein.

2. NM_007306.2: Suppressed sequence
Description NM_007306.2 was permanently suppressed because currently there is insufficient support for the transcript and the protein.

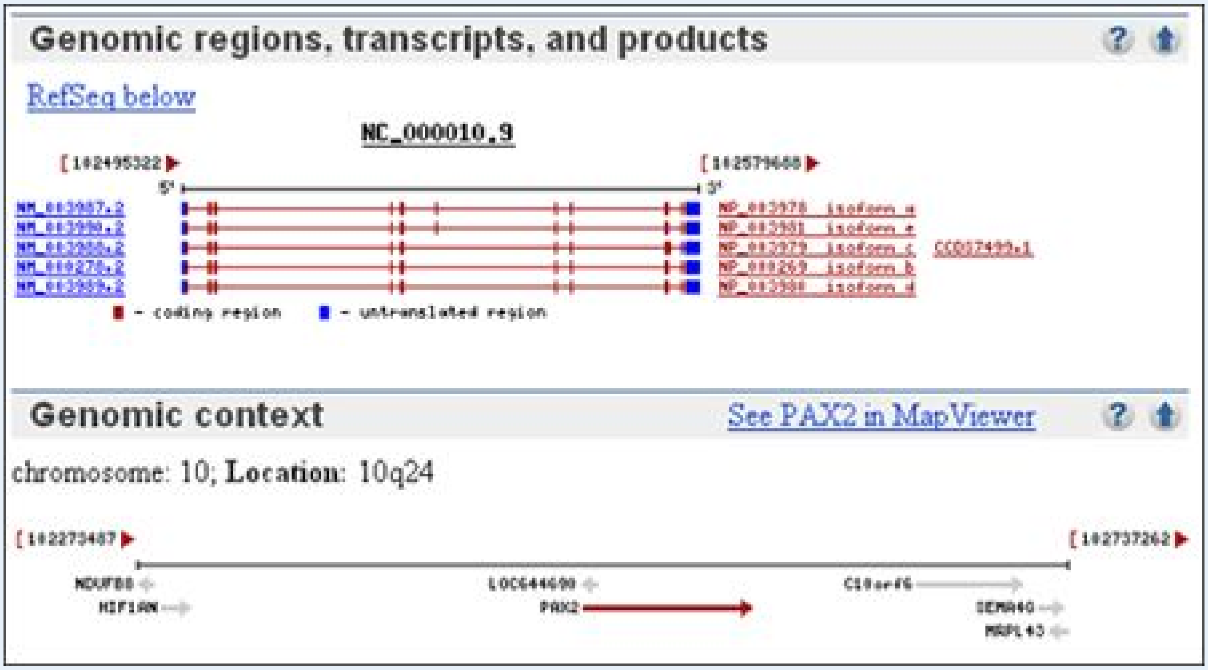
Figure 8. Representative subsection RefSeqs of Annotated Genomes**(A)** in the NCBI Reference Sequences (RefSeq) section of a Full Report display. This subsection follows RefSeqs maintained independently of Annotated Genomes (see Figure 7). It includes the accession numbers in the available Genomic assemblies, in this case the Reference assembly and the Alternative assembly from Celera with links to the GenBank and FASTA records **(B)**. In this example, the gene is annotated on the complementary strand in all genomic accession numbers **(C)**, and no model mRNA or protein accession numbers were created. Note that some RefSeqs for this gene have been suppressed and are listed under the Suppressed Reference Sequences subsection **(D)**.

### Table 2. Access to sequence information from the Gene Table display option.

| Scope | Link to use |
|---|---|
| Complete gene feature | Options from the menu opened by clicking on the genomic RefSeq accession (format NC_, NW_ NT_) at the top of the graphic display. |

| Scope | Link to use |
|---|---|
| Complete RNA | Options from the menu opened by clicking on the RNA RefSeq accession (format NM_, NR_, XM_, XR_) at the left of the graphic display. Display in GenBank format from the RNA RefSeq accession at the top of the accession-specific subsection of the table. |
| Complete protein | Options from the menu opened by clicking on the Protein RefSeq accession (format NP_, XP_) at the left of the graphic display. Display in GenBank format from the protein RefSeq accession at the top of the accession-specific subsection of the table. |
| Single intron, exon, or coding sequence (CDS) | Display in FASTA format from the range shown in the corresponding table column. |

## Box 1. Interactive View



chromosome: 10; **Location**: 10q24

## Accessing the Genomic Sequence

NC_000010.9 is the accession and version of the genomic RefSeq sequence that contains the gene. The region being shown in this accession is indicated by the integers to the left and right of the accession. The rightward-pointing arrow indicates that these mRNAs are annotated in the same orientation as the NC accession (because this is an example of a human chromosome, pter→qter). If mRNA is annotated on the reverse complement of the genomic sequence, then the phrase (shown on the reverse complement genome) is printed. RefSeq below provides a link to the RefSeq section, where there may be more information about these sequences.

Clicking on the accession NC_000010.9 brings up a menu that facilitates navigation to:

**FASTA**: The genomic sequence of the specified range, in the default orientation (not necessarily that of the annotation), in FASTA format.

**GENBANK**: The genomic sequence of the specified range, in the default orientation (not necessarily that of the annotation), in GenBank format.

Please note that whatever option you select, after you have navigated to Entrez Nucleotide, you can select any of the display options there (Graph, XML, etc.) to customize your view.

When you have used any of the above links to display the genomic sequence of the gene displayed on the record (in this case, human PAX2), you can use standard Entrez functions to:

- download the sequence in any format using the desired combination of Display and Send to options
- modify the limits of the sequence to be displayed or downloaded using the Range: from ...tooption
- add or suppress the display of optional features using the Features option

For more details about the use of Entrez features, please refer to the general Entrez Help [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp] documentation.

## Display of Products and Accessing Their Sequences

Under the line representing the genomic sequence is a description of the products of the gene. If the gene is a non-transcribed pseudogene, then only the diagram of the genomic sequence is shown.

Accession numbers listed at the left of the diagram (e.g., NM_003987.2, NM_003989.2, etc.) represent RefSeq RNAs and the versions annotated on the genomic RefSeq. Each accession.version anchors a link to retrieving sequence, formatted as described (FASTA, GENBANK) for the genomic RefSeq. If there are accession numbers listed at the right, these are for encoded proteins. For a few genomes, there may be an additional link to the Consensus CDS dataset (CCDS) [http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi], indicating that this coding region for this protein has been annotated consistently by the CCDS collaboration. In addition to supporting links to sequence, the protein accession numbers facilitate connections to protein-specific tools or displays, such as:

**BLink (BLAST Link)**: Displays the results of BLAST searches that have been done for every protein sequence in the Entrez Proteins data domain and supports multiple tools for comparing proteins.

**CDD (Conserved Domain Database)**: Enables a review of domains found in the protein, as well as links to records of other proteins with a related domain.

Each protein accession may also have a distinguishing label. In this example, these are the isoform designations, isoforms a, b, c, and d.

The labels of the RNA and protein accession numbers are color coded to coincide with the diagram of the placement of the exons. Blue indicates RNA, and in protein-coding genes, the untranslated region (UTR) of any exon. Maroon indicates protein and, therefore, is used to represent the coding region of an exon. If an intron is flanked by coding sequence, then the line connecting the exons is maroon; otherwise, it is blue.

# Finding Data Related to Entrez Gene in Other Entrez Databases

Many of the display options available from the Display drop-down menu (second gray bar at the top of the search results page) (Figure 1C) are designed to make it easier for you to retrieve information related to all or a selected set of your query results from other databases in NCBI's Entrez system. When you submit a search to Entrez Gene, the Entrez system also determines what other databases in the system contain information related to your search results. This behind-the-scenes, precomputation of relationships makes finding sets of records in other databases at NCBI that are related to your query results fast and easy. These options are based on the same infrastructure that supports the Links menus (Figure 2) available for records in Entrez Gene one at a time. The options in the display section give the added advantage of allowing display of records in another database connected to all (the default) or a selected subset (using the check boxes) of record in your result set.

## LinkOut

LinkOut [http://www.ncbi.nlm.nih.gov/entrez/linkout] provides easy access to relevant online resources outside of the Entrez system. These connections are maintained by the the external database.

## Book Links

An increasing number of Gene records are annotated specifically in books and monographs provided in Bookshelf. One example, restricted to human genes, is the GeneReviews [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=gene.TOC&depth=1] book provided in collaboration with the GeneTests group of the University of Washington. To retrieve all records in Entrez Gene in this category, try the query "gene books"[filter] from the Entrez Gene Search bar.

## Conserved Domain Links

Protein sequences are routinely compared to canonical sequences for domains in the Conserved Domain Database. Domain records connected to protein associated with records associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene in this category, try the query "gene cdd"[filter] from the Entrez Gene Search bar.

## Genome Links

Genome maintains information about chromosomes and complete genomes. Genome records associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene in this category, try the query "gene genome"[filter] from the Entrez Gene Search bar.

## GENSAT Links

GENSAT maintains images of the expression of a subset of genes expression in the central nervous system and eye of the laboratory mouse. GENSAT records associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene in this category, try the query "gene gensat"[filter] from the Entrez Gene Search bar

## Geo Profile Links

GEO maintains information from array-based experiments. Links between the databases are calculated when both GEO and Gene have computed a relationship to the same sequence record. To retrieve all records in Entrez Gene in this category, try the query "gene geo"[filter] from the Entrez Gene Search bar.

## HomoloGene Links

HomoloGene compares protein-coding genes in several key genomes to identify homologs. HomoloGene records associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene in this category, try the query "gene homologene"[filter] from the Entrez Gene Search bar.

## NIH cDNA clone links

Some of the mRNAs associated with your Entrez Gene search results are available from NIH-supported cDNA repositories. Reports of clones in the Nucleotide database associated with your Entrez Gene search results can be retrieved by using this display option.

## Nucleotide Links

Nucleotide sequences associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene with nucleotide sequence information, try the query "gene nucleotide"[filter] from the Entrez Gene Search bar.

## OMIA Links

Online Mendelian Inheritance in Animals (OMIA) maintains information about Mendelian disorders in animals. OMIA records associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene in this category, try the query "gene omia"[filter] from the Entrez Gene Search bar.

## OMIM Links

OMIM records associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene in this category, try the query "gene omim"[filter] from the Entrez Gene Search bar.

## PMC Links

Publications available as full text from PubMedCentral may include explicit references to Gene. Publications may also be connected to Gene via a PubMed ID. PubMedCentral records associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene in this category, try the query "gene pmc"[filter] from the Entrez Gene Search bar.

## Probe Links

Probe records, such as those for resequencing primers or RNAi sequences, related to your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene in this category, try the query "gene probe"[filter] from the Entrez Gene Search bar.

## Protein Links

Protein sequences associated with your Entrez Gene search results can be retrieved by using this display option. To retrieve all records in Entrez Gene with protein sequence information, try the query "gene protein"[filter] from the Entrez Gene Search bar.

## PubMed Links/PubMed (GeneRIF) LInks

PubMed citations associated with your Entrez Gene search results can be retrieved by using this display option. Those that were generated from GeneRIFs are indicated by the PubMed (GeneRIF) option. To retrieve all records in Entrez Gene with citations in PubMed, try the query "gene pubmed"[filter] from the Entrez Gene Search bar.

## SNP Links/GeneGenoType Links

Use these display options to navigate to information about variation reported in the dbSNP database for the gene records in your search results. To retrieve all records in Entrez Gene with reported variation, try the query "gene snp"[filter] from the Entrez Gene Search bar.

## Taxonomy Links

Use this display option to navigate to information about the taxonomy of the genomes in which the gene records in your search results are found.

## UniGene Links

Use this display option to navigate to information about expression and EST sequences related to the gene records in your search results. These links are calculated from nucleotide accession numbers that are common to records in both databases. To retrieve all records in Entrez Gene with additional information in UniGene, try the query "gene unigene"[filter] from the Entrez Gene Search bar.

## UniSTS Links

Use this display option to navigate to information about PCR primers associated with gene records in your search results. This gene-to-marker association is calculated from reports from external databases or by e-PCR-based matches to cDNAs or annotated genes. To retrieve all records in Entrez Gene with additional information in UniSTS, try the query "gene unists"[filter] from the Entrez Gene Search bar.

# Query Tips: How to Use the Tabs on the Search Bar

All functions of the Entrez indexing and query engine are used by Entrez Gene. This section will therefore summarize only how to use the tools in the context of the Gene database. Entrez Help [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp] provides information on how to use the tabs History [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.section.EntrezHelp.Using_Your_History], Clipboard [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.section.EntrezHelp.Details_Button__Send], and Details [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.section.EntrezHelp.Details_Button__Send]. For general information about Entrez, see Entrez Help [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp].

## Limits

- Introduction
- Examples

### Introduction to Limits

The Limits page allows you to set the context for making queries to Entrez Gene. It is accessed from the Limits tab at the lower left in the gray query bar (Figure 9A).

Limits is designed to make it easier to execute certain queries by checking boxes, rather than by writing out the text of an Entrez query. It is particularly useful if you want to retrieve genes only:

- with a value found in a single known field (Figure 9B)

- from a particular cellular source or RefSeq representation (Exclude/Include only, Figure 9C)
- represented by a particular type of RefSeq (Limit by RefSeq Status, Figure 9D)
- from a taxonomic group (Limit by Taxonomy, Figure 9E)

Remember that once you have set Limits, that setting remains through multiple queries, unless you remove the setting. A yellow banner appears below the Limits tab when Limits is turned on. Turn off Limits at any time by removing the check to the left of the word Limits in the query bar of the result page.

The Exclude section (Figure 9C) enables you to prevent certain types of genes from being included in your result set. Each check box is independent, so if you want to prevent the retrieval of genes encoded by mitochondria and by plastids, check both boxes. The NEWENTRY option refers to the GeneID used to support submission of GeneRIFs, by species, for a gene not currently in Entrez Gene. Checking All of the above prevents all in the section from being retrieved.

You canuse the Include Only option (Figure 9C) to have only certain types of genes in your result set. These are defined as:

- Genomic: Genes encoded by chromosomes or the major genomic macromolecule for the taxon.
- Mitochondria: Genes encoded by mitochondria.
- Plasmids: Genes encoded by plasmids.
- Plastids: Genes encoded by plastids.
- RefSeqs: Genes for which RefSeqs exist.
- NEWENTRY: GeneIDs used to support submission of GeneRIF [http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html]s, by species, for a gene not currently in Entrez Gene.

Additional limits are:

- Limit by RefSeq Status [http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status] (Figure 9D): To retrieve genes based on the type of RefSeq used to represent the gene.
- Limit by Taxonomy (Figure 9E): Make it easier to restrict your query by organism.

The additional limits are treated as Include Only and are hierarchical. For example, to limit your results to genes in invertebrate genomes, check Invertebrates. These selections are also treated as the Boolean operator OR, so if you want to retrieve genes from either *Danio rerio* or *Xenopus* sp., check both Danio rerio and Xenopus.

Field restriction in the Limits page is an option that allows you to retrieve records only when your query term exists in the selected field. These fields are the same as those described in more detail in the Preview/Index section.

The Limits page also allows you to restrict queries by date. The options are:

- Creation Date (Figure 9F): Retrieve records created within the range entered, or according to pre-selected ranges in the pull-down menu.

- Last Modification (Figure 9G): Retrieve records modified within the range entered, or according to preselected ranges in the pull-down menu.

## Examples Using Limits

A. To retrieve non-mitochondrially encoded NADH dehydrogenases from human, mouse, or rat, use the Limits form to:

1. Enter **nadh dehydrogenase** in the query box (case does not matter).
2. Select Gene/Protein name from the All fields drop-down menu.
3. Check Mitochondrion under Exclude.
4. Check Homo sapiens, Mus musculus, Rattus norvegicus.
5. Press Go in the search bar.

B. To retrieve *E. coli* genes related to tryptophan, use the Limits form to:

1. Check Escherichia coli.
2. Enter **tryptophan** in the query box.
3. Press Go in the search bar.

Table 3 summarizes fields, filters, and properties that are used to categorize information in Gene records. The table also provides examples of how to use these entities effectively to retrieve records. The table is alphabetized by the values in the Field Name menu.

# Preview/Index

The Preview/Index page on any Entrez database is a powerful resource to construct useful queries and to view terms that have been indexed under any field name. Table 3 in the previous section described the fields used in indexing the records and provided some representative queries using those fields. This section will:

- describe filters in general and how they can be used to find records of interest in gene
- describe the properties assigned to gene records and provide examples of how to use them

## Filter

The term *filter* is used to describe categories of records that are grouped according to their relationship either to other Entrez databases or to external resources that have submitted LinkOut connections. If the former, the filter is named according to the pattern “gene other_Entrez_database”, such as “gene protein”. If the latter, the first two letters of the filter's name are ”lo”, for LinkOut. For a comprehensive listing of filters valid for the Gene database and the number of records in each, follow these steps:

1. Click on the Preview/Index tab under the query bar.
2. Use the pull-down menu named All Fields and select Filter.
3. Click on the Index button to the right of Preview to see the filter names and the number of instances of each.

Filters are powerful tools to retrieve records of interest. For example, to retrieve all records for human genes that are associated with OMIM (i.e., have connections to OMIM) and have been annotated on the genome, use the "AND" operator with both "gene omim" and "gene nucleotide pos". Please refer to Table 4 for the current set of filters.

## Properties

Properties are assigned to gene records based on content, rather than relationship to other database records, which is the role of filters (see Filter). At present, the properties assigned to Gene records fall into these major categories:

- Type of gene: Property named as *genetype name_of_type*.
- Source of the gene: Property named as *source name_of_source*.
- Type of RefSeq provided for the gene: Property named as *srcdb refseq type_of_refseq*.
- Other

The genetype option should be self-explanatory, except perhaps for *miscrna*, *other*, and *unknown*. Names for the types of molecules encoded by genes follow the conventions [http://www.ncbi.nlm.nih.gov/projects/collab/FT] of the collaborating sequence databases (DDBJ/EMBL/GenBank); thus *miscrna* (misc_rna, miscellaneous RNA) is assigned to any gene that encodes an RNA product not included in the other specifics. The *genetype other* property is applied to loci of known type, but a specific category has not yet been applied in the Entrezgene data model (i.e., named fragile sites). The *genetype unknown* property is applied to probable genes for which the type is still under review. This category is frequently used when the defining sequence has uncertain coding propensity. **Note:** At the time of this writing, the assignment of gene records to the latter two categories continues to be refined. We appreciate your suggestions for any improvements.

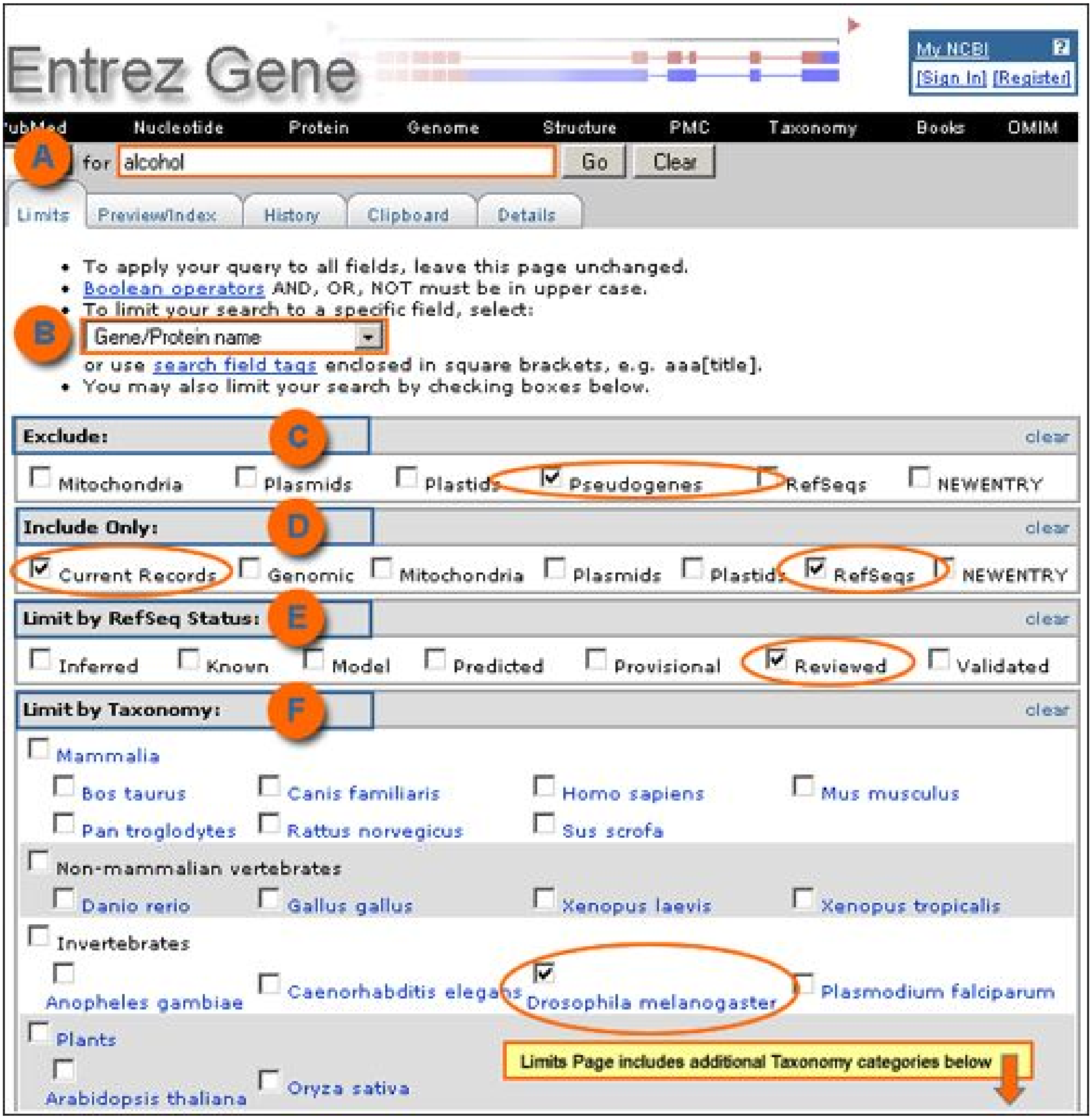
The source options should be self-explanatory, with source other being used where a specific category has not yet been applied in the Entrezgene data model.

The *srcdb refseq* categories are as enumerated by RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status] and will not be duplicated here.

The "other" properties are explained in Table 5.

## History, Clipboard, and Details

Entrez Help provides information on how to use the tabs History [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.section.EntrezHelp.Using_Your_History], Clipboard [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.section.EntrezHelp.Details_Button__Send], and Details [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.section.EntrezHelp.Details_Button__Send].



Entrez Gene

My NCBI ?
[Sign In] [Register]

ubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM

for alcohol Go Clear

Limits Preview/Index History Clipboard Details

- To apply your query to all fields, leave this page unchanged.
- Boolean operators AND, OR, NOT must be in upper case.
- To limit your search to a specific field, select:

  Gene/Protein name

  or use search field tags enclosed in square brackets, e.g. aaa[title].
- You may also limit your search by checking boxes below.

**Exclude:** clear

□ Mitochondria □ Plasmids □ Plastids ☑ Pseudogenes □ RefSeqs □ NEWENTRY

**Include Only:** clear

☑ Current Records □ Genomic □ Mitochondria □ Plasmids □ Plastids ☑ RefSeqs □ NEWENTRY

**Limit by RefSeq Status:** clear

□ Inferred □ Known □ Model □ Predicted □ Provisional ☑ Reviewed □ Validated

**Limit by Taxonomy:** clear

□ Mammalia
□ Bos taurus □ Canis familiaris □ Homo sapiens □ Mus musculus
□ Pan troglodytes □ Rattus norvegicus □ Sus scrofa

□ Non-mammalian vertebrates
□ Danio rerio □ Gallus gallus □ Xenopus laevis □ Xenopus tropicalis

□ Invertebrates
□ Anopheles gambiae □ Caenorhabditis elegans ☑ Drosophila melanogaster □ Plasmodium falciparum

□ Plants
□ Arabidopsis thaliana □ Oryza sativa

Limits Page includes additional Taxonomy categories below

Figure 9. Representative Limits Page. Shown is an example of a Limits page used to find *current records* in Entrez Gene for *Drosophila melanogaster* that are associated with *reviewed* RefSeqs, are not *pseudogenes*, and have *alcohol* in the *gene name*. **(A)** Enter alcohol in the query box. **(B)** Set the filed name to Gene/Protein name. **(C)** See Exclude and check Pseudogenes. **(D)** See Include Only and check Current Records and RefSeqs. **(E)** See Limit by RefSeq Status and check Reviewed. **(F)** In Limit by Taxonomy, check Drosophila melanogaster in the Invertebrates subsection.

## Table 3. Fields, filters, and properties in Entrez Gene.

| Field name | Definition [including field abbreviations] | Examples |
|---|---|---|
| **Name subcategory** | | |
| Disease name or phenotype of mutants | Disease or phenotype associated with the record. [DIS] | Find the genes that contribute to SCID. SCID[dis] |
| Gene name | A symbol for the gene. Includes preferred symbols, aliases, and locus tags. [SYM][SYMB][GN][GENE NAME] | Genes with a symbol starting with smt. **smt*[sym]** |
| Gene/protein name | The short or full name of the gene or any of its protein products (when applicable). | Find genes that have the word kinase in GO annotation but do not have the word kinase in the name.<br>kinase[gene ontology] NOT kinase[gene/protein name] |
| **Location subcategory** | | |
| Chromosome | Chromosome location of the gene. The value used is according to the convention of the source genome. In other words, if III is used, III but not 3 will be indexed in this field.<br>[CHRM][CHR][CHROMOSOME] | Retrieve records containing the word kinase, and the gene is located on chromosome III:**kinase AND III[chr]**<br>Retrieve records containing the words zinc and finger that are of human origin but not on chromosome 19: **zinc finger NOT 19[chr] AND "Homo sapiens"[orgn]** |
| Default map location | A map location in the units standard for the genome. For example, for human it is the cytogenetic band, for mouse it is the MGI map (centiMorgans). This is processed as a text field, so range queries are not implemented. For range queries, use Map Viewer [http://www.ncbi.nlm.nih.gov/mapview]. | Rat genes mapped to 18 q:<br>rat[orgn] AND 18q[default map location] |
| **Sequence subcategory** In Gene**: This means searching by sequence identifier, not by the sequence itself, which is managed by BLAST** [http://www.ncbi.nlm.nih.gov/BLAST]**.** | | |
| Nucleotide accession | An accession for a nucleotide sequence. [NACC] | There are instances where the same accession is applied to both nucleotide and protein sequences. To restrict an accession to nucleoide, use this field. (Accession numbers beginning with BC are not in this category.)<br>BC052629[NACC] |
| Nucleotide UID | The gi of a nucleotide sequence.[NUID][NUCL_UID][NUCLEOTIDE_UID] | Many integer identifiers have overlapping number spaces. To find the gene record that corresponds to a given nucleotide gi from gene, use this field.<br>27363473[NUID] |
| Protein accession | An accession for a protein sequence. [PACC][PROT_ACCN] | There are instances where the same accession is applied to both nucleotide and protein sequences. To restrict an accession to protein, use this field. (Accession numbers beginning with three letters are not in this category.)<br>AAH52629[PACC] |
| Protein UID | Protein gi. [PUID][PROT_UID][PROTEIN UID] | Many integer identifiers have overlapping number spaces, So to find the gene record that corresponds to a given protein gi from gene, use this field.<br>27363473[PUID] |

| Field name | Definition [including field abbreviations] | Examples |
|---|---|---|
| Nucleotide or Protein accession | A sequence accession of any type. [ACCN] | Find all the genes encoded in accession AE003828:<br>AE003828 |
| **Miscellaneous subcategory (alphabetical)** | | |
| Creation date | Date the record was created.<br>[cd][cdat][creation date] | Records containing the word xenopus created between February 5, 2004 and February 12, 2004:<br>**2004/2/5:2004/2/12[cd] AND xenopus [orgn]** |
| EC/RN number | Enzyme commission identifier for a product of the gene. Indexed without the EC prefix.<br>[ECNO][EC] | Retrieve records where proteins have an E.C. number of 1.9.3.1: **1.9.3.1[ECNO]** |
| Filter | Find records with a relationship to other data in Gene. For more examples of use of filters, see the Preview/Index section. | Retrieve records of mouse kinase genes with expression data stored in GEO:<br>**mouse[orgn] AND gene_geo[filter] AND kinase** |
| Gene Ontology | GO terms applied to this gene AND the GO identifer as the integer. The terms include the component, function, and process categories.[GO][GENE ONTOLOGY] | Rat genes with GO terms starting with “kinase signaling”<br>kinase signaling*[gene ontology] rat [orgn]<br>Any gene with the GO id of GO:0004872:<br>4872[GO] |
| LocusLink ID | The gene identifier from LocusLink. [LID]<br>[LOCUS_ID] | Retrieve the record where LocusID =2: 2[LID] |
| MIM | Identifier assigned to human genes and phenotypes by OMIM [MIM] | Retrieve records that contain the MIM number 181510: 181510[MIM] |
| Modification date | Last date the record was modified.<br>[MODDATE][MDAT][LMOD][DATE]<br>[UPDATED][MD] | Retrieve records for genes from eubacterial genomes last modified after March 10, 2004:<br>**eubacteria[orgn] AND 2004/3/10:2010/1/1[md]**<br>Retrieve records from sea urchins modified in the last 30 days:<br>**echinoidea[orgn]+AND+"last 30 days"[mdat]** |
| Property | An attribute of a Gene record based on its content.[prop][property] | Mouse records with transcript variants:<br>mouse[orgn] AND "has transcript variants"[property] |
| PubMed UID | PubMed id. [PMID] | Many integer identifiers have overlapping number spaces. To find the gene record (s) that corresponds to a paper in PubMed from Gene, use this field:<br>12477932[PMID] |
| Taxonomy ID | Identifier for the species or strain in the NCBI taxonomy database. HINT: txid{value} also works, e.g., txid9606.[TAXID][TID] | Find all records in Entrez Gene for the pig:<br>9823[taxid]<br>Alternatively:<br>txid9823 |
| Text Word | Any word in the record.[TEXT][WORD][AB][TXT] | Retrieve records that contain “32” in a record that also contains threonine, serine, and kinase: **serine AND threonine AND kinase AND 32[TEXT]** |
| UniGene cluster number | UniGene cluster including the text prefix.<br>[UNIGENE][UGEN] | Hs.2[UNIGENE] |

**Table 4. Filter sets.**

| Filter name | Definition |
|---|---|
| all | Total records, current or not |
| gene all | All current records |
| gene books | Gene records with explicit links to Entrez Books |
| gene gensat | Gene records with explicit links to Entrez GenSAT |
| gene geo | Gene records with explicit links to Entrez GEO |
| gene homologene | Gene records with explicit links to Entrez HomoloGene |
| gene nucleotide | Gene records with explicit links to Entrez Nucleotide, excluding RefSeq chromosome or contig accessions |
| gene nucleotide pos | Gene records with explicit links to Entrez Nucleotide, limited to those of RefSeq chromosome or contig accessions, and thus including position data |
| gene omim | Gene records with explicit links to Entrez OMIM, and thus includes links to both disease and “gene” records in OMIM |
| gene protein | Gene records with explicit links to Entrez Protein, and thus includes links to GenPept and SwissProt accessions |
| gene pubmed | Gene records with explicit links to Entrez PubMed |
| gene snp | Gene records with explicit links to Entrez dbSNP, and thus supports finding gene variation information available in dbSNP |
| gene taxonomy | Gene records with explicit links to Entrez Taxonomy |
| gene unigene | Gene records with explicit links to Entrez UniGene |
| gene unists | Gene records with explicit links to Entrez UniSTS (marker data) |

**Table 5. Other properties.**

| Property name | Explanation |
|---|---|
| alive | The record is current and primary (i.e., not secondary or discontinued). The term *secondary* is applied to any record that has been merged into another. This occurs most often when multiple genes are defined based on incomplete data, and these are later discovered to be parts of the same gene. One gene record then becomes secondary to the other. |
| GeneRIF | A record having one or more GeneRIF annotations attached. |
| has ccds | A gene that encodes a protein sequence that is a member of a Consensus CDS (CCDS). See http://www.ncbi.nlm.nih.gov/projects/CCDS/. |
| has transcript variants | A record having two or more associated RefSeq transcripts, i.e., splice variants. **Note:** This is limited to RefSeq annotation and should **not** be used to identify all genes exhibiting alternative splicing, promoter usage, and/or polyadenylation signals. |

# Constructing Powerful Queries

Constructing queries based on free text, filters, and properties can be quite powerful in retrieving records of interest from Gene. Table 6 summarizes some of these approaches by describing:

- Scope: The intent of a query.

- Query: How to construct a query that meets that intent.
- Notes: How usage of Gene to retrieve these data may compare to other gene-related resources, namely HomoloGene, Map Viewer, or UniGene.

Although these examples use field restriction (see Table 3 for the comprehensive list of fields, filters, and properties in Entrez Gene), free text can also be submitted. Entrez Gene then weights the retrievals based on the field in which a result was found. For example, if your query matches a gene symbol in one record and arbitrary text in another, the record where the match is on the symbol will be displayed before the other in the results.

**Table 6. Constructing queries.**

| Scope | Query | Notes |
|---|---|---|
| Find genes mapped to *Arabidopsis thaliana* chromosome 3 that have orthologs reported in HomoloGene | arabidopsis thaliana[orgn] AND 3[chr] AND gene_homologene[filter] | [orgn] is used to restrict “Arabidopsis thaliana” to the organism field. That restriction could also be set by checking **Arabidopsis thaliana** on the Limits form.<br>[chr] is used to restrict ”3” to the chromosome field.<br>gene homologene[filter] is used to restrict records to those processed by HomoloGene. This query is not currently able to be processed by MapViewer, because the relationship to HomoloGene is not processed for indexing at present, nor by HomoloGene, because the chromosome data are not captured in HomoloGene. |
| Find genes also being processed by OMIM but for which there is not currently a RefSeq of the type "known" | gene omim[filter] NOT srcdb refseq known [prop] | gene omim[filter] is used to find all Gene records with relationships to OMIM. srcdb refseq known[prop] is used (as the Boolean NOT) to find all such records that do not have RefSeqs of the accession format NM_000000, NG_000000, or NR_000000. |
| Find genes from genomes other than mammals that are classified by the GO consortium to have some relationship to the cytoskeleton | cytoskeleton[go] NOT mammalia[orgn] | [go] is used to restrict to the field “Genome Ontology”.<br>[orgn] is used to restrict (as the Boolean NOT) to species not classified as mammals. Queries based on GO terms are not supported in either Map Viewer or HomoloGene. Please note that Gene does not recapitulate tree-based searching for GO annotation; this retrieval is based solely on the existence of the word in any GO category. Links are provided to the GO website to support more specific searches. |

# Tips for Programmers

## The Gene Data Model and DTD

The data model for Entrez Gene is documented in the Entrezgene specification [http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/entrezgene/entrezgene.asn]. It combines several definitions used by other NCBI databases, such as seqfeat [http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/seqfeat/seqfeat.asn], but also establishes definitions specific to Entrez Gene. Of special note is the Gene-commentary, which is used to represent many descriptors of genes. Each Gene-commentary is defined by type and supports specific representation of such elements as sequence database accession numbers (accession, version), citations (refs), external or internal resources defining the data (source), and position information. Heading, label, and text are used for general data, with the choice influenced by display in the Entrez Gene viewers.

The DTD for Gene is available from NCBI's DTD directory [http://www.ncbi.nlm.nih.gov/dtd] and is called NCBI Gene.dtd [http://www.ncbi.nlm.nih.gov/dtd/NCBI_Gene.dtd].

## Entrez Programming Utilities and Gene

The full power of Entrez Programming Utilities (e-Utils) can be used to extract information from Entrez Gene programmatically. The basic strategy is to identify the query that will return the desired records and then submit that query via ESearch. The GeneIDs identified by that search can then be submitted to another function, such as ESummary or EFetch. Examples for Gene are provided on the FAQ page. The FTP site [ftp.ncbi.nlm.nih.gov/gene/tools] contains a sample perl script that uses ESearch and ESummary.

## Entrez Gene FTP Site

The FTP [ftp.ncbi.nlm.nih.gov/gene] site for Entrez Gene ( README [ftp.ncbi.nlm.nih.gov/gene/README]) has three subdirectories: DATA [ftp.ncbi.nlm.nih.gov/gene/DATA], GeneRIF [ftp.ncbi.nlm.nih.gov/gene/GeneRIF], and Tools [ftp.ncbi.nlm.nih.gov/gene/tools].

DATA contains files that provide key attributes of genes, including:

- all associated accession numbers, including RefSeqs (gene2accession.gz)
- associated RefSeq accession numbers (gene2refseq.gz)
- citations (gene2pubmed.gz)
- nomenclature, ID, and map data (gene_info.gz)
- MIM numbers (mim2gene)
- UniGene clusters (gene2unigene)
- GO terms (gene2go)

Details of the construction of these files are reported in the ( README [ftp.ncbi.nlm.nih.gov/gene/README]) file.

DATA also contains the ASN_BINARY [ftp.ncbi.nlm.nih.gov/gene/DATA/ASN_BINARY] subdirectory. This path contains both a comprehensive extraction from Gene (All_Data.gz), several subsets categorized by source (Organelles, Plasmids), and subdirectories grouped broadly by taxonomy. The format of these extractions is compressed binary ASN.1. The program gene2xml [ftp.ncbi.nlm.nih.gov/asn1-converters/by_program] is available to convert these files to XML or ASN.1 text.

GeneRIF contains files that provide supplemental information about gene functions, either from the GeneRIF [http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html] pipeline (generifs_basic.gz) or the HIV-1, Human Protein Interaction Database [http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions] (hiv_interactions.gz). The tab-delimited files are not subdivided by species of interest. All files except the file reporting GeneID/PubMedID relationships (gene2pubmed.gz) have a column with the ID from the NCBI Taxonomy database to facilitate the extraction of a subset of the data from the file by species.

Gene_tools [ftp.ncbi.nlm.nih.gov/gene/tools/README] provides or points to programs and scripts to mine data from Entrez Gene. Of particular interest is gene2xml [ftp.ncbi.nlm.nih.gov/asn1-converters/by_program], which can be used to convert the binary ASN.1 in the ASN_BINARY directory to XML or to ASN.1 in text format ( README [ftp.ncbi.nlm.nih.gov/gene/tools/README]).

## Connecting Users of Entrez Gene to Your Website

Entrez Gene can serve as a gateway to information on your website served from your local database. Users of Entrez Gene will discover your website if you participate in our LinkOut [http://www.ncbi.nlm.nih.gov/entrez/linkout] system and become a LinkOut provider. Any Entrez database will support LinkOut. Linkout Help’s Information for Other Resource Providers [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helplinkout.chapter.nonbib] explains the details of this opportunity.

There are many benefits to becoming a LinkOut provider. If you want access to your database to be apparent from Entrez Gene, you can control the description of your resource, the update cycle, and the icon to anchor links to your site. In other words, you do not have to wait for NCBI staff to go to your site to obtain and process information and match to Gene records. You know your site best—you can identify which records are related to Gene records and provide the most accurate and informative URL to connect that Gene record to your site. If you already provide LinkOuts to other Entrez databases, such as Nucleotide or Protein, you do not have to re-register as a provider; you need only notify the LinkOut [mailto:] staff and start to submit a new resource file.

With the implementation of My NCBI [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpmyncbi.chapter.MyNCBI], it is even more advantageous to become a LinkOut provider. One of the options registered users of My NCBI can select is to display the icons for any LinkOut provider at the top of a record. The presence of your familiar logo would invite users of Entrez Gene to go to your site.

# Historical Information about LocusLink

Earlier versions of the help documentation for Entrez Gene included tips for the user accustomed to using LocusLink (discontinued in 2005). This help document has consolidated portions of that text, namely correspondence of link names (Appendix 1) and mapping of data from the LLtmpl format to Entrez Gene's ASN.1 (Appendix 2 and Appendix 3). Some of the steps of the transition process are also documented here.

### Appendix 1. Correspondence between LocusLink letters and Entrez Links.

| Letter in LocusLink | Link name | Scope |
|---|---|---|
| P | PubMed | All citations, including those established via GeneRIFs |
| O | OMIM | The OMIM records based on the gene or the phenotype (human only) |
| R | | Not implemented separately; included in Nucleotide |
| G | Nucleotide | A subset of nucleotide links. ESTs are excluded unless the gene is defined by an EST; unannotated, high-throughput genomic (HTG) sequences are also excluded unless used as a source for a RefSeq. |
| P | Protein | A subset of protein sequences encoded by the gene |
| H | HomoloGene | HomoloGene links based on a shared GeneID |
| U | UniGene | UniGene links based on shared nucleotide sequences |
| V | SNP | Links to dbSNP for all variations related to the GeneID, variant by variant |
| | GeneView in dbSNP | A display of all variations related to the GeneID, by placement in the gene. This is the view directly corresponding to the V link |

Note: Gene supports many more links from the menu at the right of the Summary and Brief views than than the color-coded letter icons that LocusLink provided. For Gene, all that is required is a connection between Gene and another Entrez database. Thus, there may be links to Books, GEO, UniSTS, Taxonomy, etc.

## Appendix 2. Relationship between LL_tmpl and Entrez Gene.

| LocusLink | Gene | Comments |
|---|---|---|
| Table of Contents | Not retained | |
| Alphabetic lists | Not retained | |
| Gene diagram | Transcripts and Products | Gene adds the function of Genomic context to allow a quick view of nearby genes and links to their report pages. |
| Link to Evidence Viewer from Gene diagram | Evidence Viewer link in Links menu | The option to first see only the diagram of the alignment is not retained. |
| Button Links | Links menu | On the Gene Graphic/Default display, the number of links may be greater than in LocusLink. |
| Title bar with links to nomenclature source | Initial text, with link to nomenclature source via LocusTag | Links from LocusTag values may connect to an external database where official nomenclature has not been assigned. |
| **Overview Section** | | |
| RefSeq Summary | Summary | Not changed. |
| Locus type | Combination of Gene type and evidence type (under development) | The text values are not equivalent. LocusLink's Locus type values are being subdivided into Gene type and Evidence type categories. |
| Protein names | General protein information | Not changed. |
| Alternate symbols | Gene aliases | Not changed. |
| **Relationships Sections** | | |
| Homology data | Links menu; HomoloGene | What is printed in LocusLink is still printed in Gene. |
| Related models | Related | Not changed; limited to genomes being annotated by NCBI's pipeline. |
| **Function Section** | | |
| GeneRIFs | GeneRIFs | Not in a function section; indented under Bibliography. |
| GO annotation | General gene information: GeneOntology | Organization changed, but not content. |
| Phenotype | General gene information: Phenotypes | Organization changed, but not content. |
| **Map Section** | | |
| Chromosome | Genomic context | Not changed. |
| Associated markers | General gene information: Sequence Tagged Site (Markers) | Entrez Gene added display of alternate marker names. |
| **Sequence Section** | | |
| **RefSeq Subsection** | | |
| Category | RefSeq status | Not changed. |
| GenBank source | Source sequence | Not changed. |
| Domain matches | Domains | CDD link also attached to the protein accession in **Transcripts and Products** and in the Links menu. |
| BL (BLink) | BLink link attached to the protein accession in the Transcripts and Products section. | The function was not changed, but the placement and visibility are different. |
| Variant name | After the protein accession in the Transcripts and Products section. | Content not changed. |
| **Annotation Subsection** | | |
| Genomic contig | Transcripts and Products | Function retained. |
| **gb**: Link to gene-specific subsequence | GENBANK view from source NC, NT, or NW accession | Function retained as GENBANK option from the genomic accession-based menu. |
| **sv:** Link to graphic display of gene-specific subsequence | GRAPHICS view from source NC, NT, or NW accession | Function retained as GRAPHIC option from the genomic accession-based menu. |
| **mv**: Map Viewer | Map Viewer in Links menu | Function retained. |
| **ev**: Evidence viewer | Evidence Viewer in Links menu | Function retained. |
| **mm**: ModelMaker | ModelMaker in Links menu | Function retained. |

| LocusLink | Gene | Comments |
|---|---|---|
| Strain or haplotype | Not retained. | |
| **Related Sequences Subsection** | | |
| Accessions, type, and strain data | Related Sequences | Content not changed. |
| BL (BLink) from protein accessions | Not retained. | |
| Additional Links | Additional Links | |

## Appendix 3. Entrezgene equivalents.

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| >>[numeric] record separator; the number equals the LocusID<br>**Note:** LocusID = geneid for those records also public in LocusLink | Record set closed by bracket matching Entrezgene ::= **{** |
| LOCUSID [numeric] [unique] [required] the unique integer id for a locus<br>**Note:** LocusID = geneid for those records also public in LocusLink | geneid<br>Entrezgene ::= {<br>track-info { **geneid** 2, |
| CURRENT_LOCUSID: [numeric] [unique] [optional] If a LocusID has been merged with another, the current LOCUSID, corresponding to the value on the previous LOCUSID line, is provided here. | current-id<br>In this example, geneid 217346 is secondary to geneid 193217.<br>Entrezgene ::= {<br>track-info {<br>geneid 217346,<br>status secondary,<br>**current-id** {<br>{<br>db "LocusID",<br>tag id 193217<br>},<br>{<br>db "GeneID",<br>tag id 193217<br>} }, |
| LOCUS_CONFIRMED: [alphanumeric][yes\|no] The LOCUSID has been assigned to a confirmed locus and can be treated as an identifier that will be tracked. | No direct equivalent at present |
| LOCUS_TYPE: [alphanumeric] description of the type of locus | type [http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/entrezgene/entrezgene.asn#L34]<br>**type** protein-coding,<br>and<br>evidence (under development) |
| ORGANISM: [alphanumeric] [unique] [required] source species (Homo sapiens, Rattus norvegicus, etc.), based on NCBI's Taxonomy | source→taxname<br>source {<br>genome genomic,<br>origin natural,<br>org {<br>**taxname** "Homo sapiens",<br>common "human",<br>db {<br>{<br>db "taxon",<br>tag id 9606<br>}<br>},<br>syn {<br>"man"<br>},<br>orgname {<br>name binomial {<br>genus "Homo",<br>species "sapiens"<br>},<br>lineage "Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo",<br>gcode 1, |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| | mgcode 2,<br>div "PRI"<br>} }, |
| RELL: [set][optional][alphanumeric][multiple] description\|id\|id type\|print representation[/set] brief text summarizing the relationship, the other id, the type of id, and the display for that second id. At present these id types of are 2 classes: l for locus_id, n for nucleotide accession official/default symbol for the other locus being described. The function of reporting the other GeneID is not retained. Consider HomoloGene's FTP site for these data. | Gene/Gene relationship of Homology<br>homology {<br>{<br>type comment,<br>heading "Mouse",<br>label "A2m",<br>text "6 62.00 cM",<br>source {<br>{<br>src {<br>db "HomoloGene",<br>tag str "A2m"<br>},<br>anchor "A2m",<br>url "/Homology/view.cgi?map=ncbi_mgd&tax_id=10090&chr=6&symbol=A2m"<br>}<br>}<br>} }, |
| STATUS: [alphanumeric] [optional] (only if a reference sequence exists) [REVIEWED\|PROVISIONAL\|PREDICTED\|MODEL\|INFERRED] type of reference sequence record | Gene-commentary of **type** comment with **heading** ="RefSeq Status" and **label** of the appropriate value<br>{<br>**type** comment,<br>**heading** "RefSeq Status",<br>**label** "REVIEWED"<br>}, |
| NC: the accession for chromosome RefSeq records [alphanumeric] [optional] (only if a reference sequence exists) the RefSeq accession for a genomic record, followed by the gi and strain, if applicable. | **locus→accession**<br>**locus** {<br>{<br>type genomic,<br>label "tat",<br>**accession** "NC_001802",<br>seqs {<br>int {<br>from 5376,<br>to 7969,<br>strand plus,<br>id **gi** 9629357<br>} }, |
| NR: The RefSeq accession for a non-messenger RNA.<br>NM: The RefSeq accession for a mRNA record [alphanumeric] [optional] (only if a mRNA reference sequence exists) the accession for the mRNA, followed by the gi and the strain, if applicable<br>NP: The RefSeq accession for a protein record [alphanumeric] [optional] (only if a reference sequence exists) the RefSeq accession number for a protein record, followed by the PID for that protein<br>XR: [alphanumeric][optional] (only if a model exists) the RefSeq accession of a model RNA, not associated with a protein product<br>XM: [alphanumeric] [optional] (only if a model exists) the accession for the mRNA, followed by the gi and the strain, if applicable<br>XP: The RefSeq accession for a model protein record [alphanumeric] [optional] (only if an XM exists) the RefSeq accession of a model protein, followed by the PID for that protein | May be in two places:<br>if annotated on a genomic RefSeq, then in **locus**, **products**, **type**="...", **accession**<br>always in comments, type **comment**, **products**, **type** ..., **heading** "...Sequence", **accession**<br>**type** comment,<br>**heading** "NCBI Reference Sequences (RefSeq)",<br>**products** {<br>{<br>**type** mRNA,<br>**heading** "mRNA Sequence",<br>**accession** "NM_000014",<br>**version** 3,<br>source {<br>{<br>src { |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| | db "Nucleotide",<br>tag str "6226959"<br>},<br>anchor "NM_000014"<br>}<br>},<br>seqs {<br>whole **gi** 6226959<br>},<br>**products** {<br>{<br>**type** peptide,<br>**heading** "Product",<br>**accession** "NP_000005",<br>**version** 3,<br>source {<br>{<br>src {<br>db "Protein",<br>tag id 4557225<br>},<br>anchor "NP_000005",<br>post-text "alpha 2 macroglobulin precursor"<br>}<br>},<br>seqs {<br>whole **gi** 4557225 }, |
| NG: The RefSeq accession for genomic region (nucleotide) records | Gene-commentary only, heading "NCBI Reference Sequences (RefSeq)"<br>{<br>type comment,<br>heading "NCBI Reference Sequences (RefSeq)",<br>comment {<br>{<br>type genomic,<br>heading "Reference",<br>accession "NG_002315",<br>version 1,<br>source {<br>{<br>src {<br>db "Nucleotide",<br>tag id 24047158<br>},<br>anchor "NG_002315"<br>}<br>},<br>seqs {<br>int {<br>from 1,<br>to 652,<br>strand plus,<br>id gi 24047158<br>}<br>}<br>}<br>} }, |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| PRODUCT: [alphanumeric] [optional] (only if a reference sequence exists) the name of the product of this transcript | Provided as **post-text** in the Gene-commentary for the protein accession<br>products {<br>{<br>type peptide,<br>heading "Product",<br>accession "NP_000005",<br>version 3,<br>source {<br>{<br>src {<br>db "Protein",<br>tag id 4557225<br>},<br>anchor "NP_000005",<br>post-text "alpha 2 macroglobulin precursor"<br>}<br>}, |
| TRANSVAR: [alphanumeric] [optional] (only if a reference sequence exists) a variant-specific description | Gene-commentary, of **type** comment, where **heading** "Transcriptional Variant". Within the Reference Sequences Gene-commentary, indented under the RNA product.<br>comment {<br>{<br>**type** comment,<br>**heading** "Transcriptional Variant",<br>comment {<br>{<br>type comment,<br>text "Transcript Variant: This variant (PAX3A) includes an alternate segment in the coding region, which causes a frameshift, and lacks<br>several segments in the 3' coding region, compared to variant PAX3. The<br>resulting protein (isoform PAX3a) has a shorter and distinct C-terminus,<br>compared to isoform PAX3. Isoform PAX3a lacks the paired-type homeodomain."<br>}<br>} }, |
| ASSEMBLY: [alphanumeric] [optional][multiple] (only if a reference sequence exists)[/SET] | Gene-commentary, of **type** other, where **heading** "Source Sequence". Within the Reference Sequences Gene-commentary, indented under the RNA product.<br>{<br>**type** other,<br>**heading** "Source Sequence",<br>source {<br>{<br>src {<br>db "Nucleotide",<br>tag str "AJ007392,S69369"<br>},<br>anchor "AJ007392,S69369"<br>}<br>}, |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| CONTIG: [SET][alphanumeric][optional][multiple] the accession.version of the RefSeq contig, the nucleotide gi, the strain, the position of the gene (from\|to\|orientation), the chromosome, and an indicator of whether this is on the reference assembly or a strain\|haplotype | |
| XG: [alphanumeric][optional] (only if an NG accession was used in the annotation process to define position of features on the contig) NG accession, nucleotide gi, strain [SET] | The function of indicating whether an NG accession was used in NCBI's annotation process is not currently retained. The NG accessions are, however, included in the Reference Sequence section. |
| EVID: [alphanumeric] [optional] (only if a model exists) text summary of the evidence for this model | The function reporting the evidence supporting an annotated gene or RNA feature is not currently retained. |
| CDD: [alphanumeric][multiple][optional] name\|key\|score\|e_value\|bit_score [/SET] [/SET] | Gene reports domain content; position of these domains is part of the annotation of the RefSeq protein. The domain information is included as a gene-commentatary of type other, with the heading **Domains** on a gene-commentary of type peptide. The e-value is not reported.<br>comment {<br>{<br>type other,<br>heading "**Domains**",<br>comment {<br>{<br>type other,<br>source {<br>{<br>src {<br>db "CDD",<br>tag id 5952<br>},<br>anchor "pfam00207: Alpha-2-macroglobulin family",<br>post-text "score:2365"<br>}<br>} }, |
| ACCNUM: GenBank nucleotide accession used related to the RefSeq record [SET][alphanumeric] [optional] [multiple] nucleotide sequence accession number (no version), nucleotide gi, strain (if applicable), 5' end of the gene in the sequence, 3' end of the gene in the sequence one accession number per line | The data previously reported as ACCNUM, TYPE, and PROT are now reported as a set of gene-commentaries starting with one of type **comment** with the heading **Related Sequences**. The nucleotide sequence information is reported under **products**, as a gene-commentary of type **mRNA**. If that nucleotide sequence has an associated accession for one or more protein products, those data are reported under **products** as type **peptide**. Accession, version, and gi are provided. The function of reporting the position coordinates if there is no protein product is not currently retained.<br>**Example of an mRNA, its encoded protein, and strain of origin**<br>type comment,<br>heading "**Related Sequences**",<br>**products** {<br>{<br>type **mRNA**,<br>heading "**mRNA**",<br>accession "AY185125",<br>version 1,<br>source {<br>{<br>src {<br>db "Nucleotide",<br>tag id 27966960 |
| TYPE: [e\|m\|g] refers to type of nucleotide sequence: e=EST m=mRNA g=genomic | |
| PROT: [SET][multiple][optional] A potentially repeating set of two values: accession and identifier (PID value) for the coding region or regions annotated on the associated nucleotide record, one line for each accession If no data are available, na is supplied. The delimiter is \|. [/SET][/SET] | |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| | },<br>anchor "AY185125"<br>}<br>},<br>seqs {<br>whole gi 27966960<br>},<br>**products** {<br>{<br>type **peptide**,<br>accession "AAO25741",<br>version 1,<br>source {<br>{<br>src {<br>db "Protein",<br>tag id 27966961<br>},<br>anchor "AAO25741"<br>}<br>},<br>seqs {<br>whole gi 27966961<br>}<br>}<br>},<br>comment {<br>{<br>type other,<br>label "**Strain**",<br>text "C57BL/6"<br>}<br>} }, |
| [OFFICIAL\|PREFERRED]_SYMBOL: [alphanumeric] [unique] [required] the symbol used for gene reports OFFICIAL: validated by the appropriate nomenclature committee PREFERRED: interim option selected for display na is used for models without evidence<br>[OFFICIAL\|PREFERRED]_GENE_NAME: [alphanumeric] [unique] [required (but may be null)] the gene description used for gene reports OFFICIAL: validated by the appropriate nomenclature committee PREFERRED: interim selected for display [**Note:** If the symbol is official, the gene_name will be official. No record will have both official AND interim nomenclature. | The preferred symbol and preferred name are reported as gene→locus and gene→desc, respectively. Whether or not these are official is not explicitly represented. If there is a value for locus-tag, the resource associated with that locus-tag should be used to determine if the names are official or interim. If locus is not supplied, however, it indicates no official symbol has been identified.<br>**A record with an official symbol and name.**<br>gene {<br>**locus** "A2m",<br>**desc** "alpha-2-macroglobulin",<br>...<br>**A record with no identified official symbol or name.**<br>gene {<br>**desc** "spongiotrophoblast specific protein",<br>maploc "17p14",<br>db {<br>{<br>db "LocusID",<br>tag id 64509<br>}<br>},<br>syn {<br>"Tpbp"<br>}, |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| | locus-tag "RGD:621454" }, |
| PREFERRED_PRODUCT: [alphanumeric] [unique] [optional] the name of the product used in the RefSeq record | The name of any RefSeq protein product is reported as part of the protein's gene-commentary, as **post-text**.<br>type peptide,<br>heading "Product",<br>accession "NP_057236",<br>version 2,<br>source {<br>{<br>src {<br>db "Protein",<br>tag id 7706625<br>},<br>anchor "NP_057236",<br>**post-text** "retinoic acid receptor, beta isoform 2"<br>}<br>},<br>seqs {<br>whole gi 7706625<br>}, |
| ALIAS_SYMBOL: [alphanumeric][multiple] other symbols associated with this gene | All aliases are listed as synonyms gene→**syn**.<br>**syn** {<br>"HAP",<br>"RRB2",<br>"NR1B2" }, |
| ALIAS_PROT: [alphanumeric][multiple] other protein names associated with this gene | All protein names are enumerated as **prot→name**.<br>**prot** {<br>**name** {<br>"retinoic acid receptor, beta",<br>"RAR-epsilon",<br>"RAR, beta form",<br>"HBV-activated protein",<br>"retinoic acid receptor beta 2",<br>"retinoic acid receptor beta 4",<br>"hepatitis B virus activated protein",<br>"retinoic acid receptor, beta polypeptide"<br>} }, |
| REL2: [set][optional][alphanumeric][multiple] LocusID of the interacting protein\| RefSeq accession of the interacting protein\| name of the interacting protein\| keyword for the type of interaction\| accession of the RefSeq protein associated with this locus\| name of the RefSeq protein at this locus\| a description of the interaction\| PubMed id(s) describing the interaction [/set] | Not yet implemented. |
| PHENOTYPE: [SET][alphanumeric][multiple] a phenotype associated with a mutation in this gene<br>PHENOTYPE_ID: [/SET] an ID used for this phenotype. For humans, this is the MIM number | Descriptions of phenotypes associated with a gene are reported in gene-commentaries of type **comment** with the heading **Phenotypes**. The name of any phenotype is provide as **text**, and the source of that name, and its identifier there, are reported as database cross-references **source**.<br>type comment,<br>heading "**Phenotypes**",<br>comment {<br>{<br>type comment,<br>**text** "Cystic fibrosis",<br>**source** {<br>{<br>src {<br>db "MIM", |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| | tag id 219700<br>},<br>anchor "MIM: 219700"<br>}<br>}<br>},<br>{<br>type comment,<br>**text** "Pancreatitis, idiopathic",<br>**source** {<br>{<br>src {<br>db "MIM",<br>tag id 602421<br>},<br>anchor "MIM: 602421"<br>}<br>}<br>},<br>{<br>type comment,<br>**text** "Sweat chloride elevation without CF",<br>**source** {<br>{<br>src {<br>db "MIM",<br>tag id 602421<br>},<br>anchor "MIM: 602421"<br>}<br>}<br>}<br>}<br>}, |
| SUMMARY: [alphanumeric][optional] a summary description of the gene, its products, its significance, and mutant phenotypes | This is optional text, represented in the ASN.1 as 'summary', after the 'gene' and 'prot' text and before 'location'. |
| UNIGENE: [alphanumeric][multiple] UniGene cluster id(s) associated with this gene | UniGene cluster designations are reported as a gene-commentary of type comment and text **UniGene** within a gene-commentary of type comment and heading **Additional Links**. The cluster designation is provided both as a db_xref and as an **anchor**.<br>{<br>type comment,<br>heading "**Additional Links**",<br>comment {<br>{<br>type comment,<br>**text** "UniGene",<br>**source** {<br>{<br>src {<br>db "UniGene",<br>tag str "**Hs.411882**"<br>},<br>anchor "**Hs.411882**",<br>url "/UniGene/clust.cgi?ORG=Hs&CID=411882"<br>}<br>} |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| | }, |
| OMIM: [numeric][optional][multiple] MIM number | MIM numbers are reported as a gene-commentary of type comment and text **MIM** within a gene-commentary of type comment and heading **Additional Links**. The MIM number is provided both as a db_xref and as an **anchor**.<br>{<br>type comment,<br>heading "**Additional Links**",<br>...<br>comment {<br>type comment,<br>text "MIM",<br>**source** {<br>{<br>src {<br>db "MIM",<br>tag str "602421"<br>},<br>**anchor** "602421"<br>}<br>}<br>}, |
| CHR: [alphanumeric][optional][multiple] the chromosome assignment | Chromosome is represented according to the NCBI-BioSource standard, namely as **source→subtype**.<br>source {<br>genome genomic,<br>origin natural,<br>org {<br>taxname "Homo sapiens",<br>common "human",<br>...<br>},<br>**subtype** {<br>{<br>**subtype** chromosome,<br>name "7" ... |
| MAP: [alphanumeric][optional][multiple] One line, consisting of a repeating set of 3 data elements, each element separated by \| the first element is the location; the second is the source (as a URL when appropriate), and the third element is the type of map information (G = genetic, C=cytogenetc) | Map data are stored under **location**, with the units of the location being reported as method map-type.<br>**An example of a cytogenetic map location.**<br>location {<br>{<br>display-str "7q31.2",<br>method map-type cyto<br>} }, |
| STS: set of STS markers [SET][alphanumeric][optional][multiple] multiline set, one marker per line marker name\|chromosome\|sts_id\|D segment\|seq_known\|evidence[/SET] evidence types are <currently either epcr, or PubMed id(s) | Markers are reported as gene-commentaries of type comment under the heading **Sequence Tagged Site (Markers)**. The UniSTS id is the value of **tag id**, the preferred name is **anchor**, and evidence is **post-text**. The function of reporting the chromosome to which the marker has been mapped is not retained. The function of enumerating all marker aliases has, however, been added.<br>{<br>type comment,<br>heading "Sequence Tagged Site (Markers)",<br>comment {<br>{<br>type comment,<br>source {<br>{ |

| LL_tmpl description | Entrezgene equivalent |
|---|---|
| | src {<br>db "UniSTS",<br>tag id 12967<br>},<br>anchor "D7S2742",<br>post-text "(e-PCR)"<br>}<br>},<br>comment {<br>{<br>type other,<br>label "Alternate name",<br>text "G00-674-897"<br>},<br>{<br>type other,<br>label "Alternate name",<br>text "G11318"<br>},<br>{<br>type other,<br>label "Alternate name",<br>text "G13271"<br>}, ... |
| COMP: Set of comparative map links [alphanumeric][optional] [multiple] c_tax_id\|c_symbol\|c_chromosome\|c_position\| c_locus_id\| q_chromosome\|symbol of the current gene\| map_name[/SET] the tax_id of the homolog, the symbol of the homolog, the homologous chromosome, the homologous position, the locus_id of the homolog, the chromosome of the source record, the map name | This function is not likely to be retained. |
| ECNUM: [alphanumeric][optional][multiple] | Enzyme Commission numbers (EC) are reported as gene-commentaries of type **property** and label **EC**. The EC number is reported as **text**.<br>comment {<br>{<br>type **property**,<br>label "**EC**",<br>**text** "1.2.3.1"<br>} } |
| BUTTON: [SET][alphanumeric][optional] an web resource accessed by a button, as well as or in addition to text<br>LINK: [/SET][alphanumeric the url underlying the button (note: if there are variation data for this locus at NCBI, the line "BUTTON: snp.gif" will be present)<br>DB_DESCR: [SET][alphanumeric][optional][multiple] The name of an external website with more information about this locus<br>DB_LINK: [/SET][alphanumeric] the URL | If retained, then as a sequence of Gene-commentary.<br>**type** comment,<br>**heading** "Additional Links",<br>**comment** {<br>{<br>**type** comment,<br>text "UniGene",<br>xtra-properties {<br>{<br>tag "UNIGENE",<br>value "Hs.74561"<br>}<br>},<br>source {<br>{<br>src {<br>db "UniGene",<br>tag str "Hs.74561" |

<table>
<tr><th>LL_tmpl description</th><th>Entrezgene equivalent</th></tr>
<tr><td></td><td>},<br>anchor "Hs.74561",<br>url "/UniGene/clust.cgi?ORG=Hs&CID=74561"<br>}<br>} },</td></tr>
<tr><td>PMID: [numeric][multiple] a subset of publications associated with this locus with the link being the PubMed unique identifier comma separated</td><td>comments, type <b>comment</b>, <b>refs</b> where the value is <b>pmid</b><br>{<br><b>type</b> comment,<br><b>refs</b> {<br>pmid 14637088,<br>pmid 14506912,<br>pmid 1370808,<br>pmid 1281457<br>} },</td></tr>
<tr><td>GRIF: [SET][alphanumeric][optional][multiple][/SET] PubMed unique identifier|comment</td><td>Under development.</td></tr>
<tr><td>GO: [SET][alphanumeric][optional][/SET] category of term|the term itself|evidence code|GO identifier| source of annotation|PubMed id(s)</td><td>Gene Ontology annotation is reported as <b>properties</b> in gene-commentaries with the heading <b>GeneOntology</b>. The provider of the annotation is indicated under <b>source</b>. The <b>Function</b>, <b>Process</b>, and <b>Component</b> categories are headed by <b>label</b>. The term is provided as <b>anchor</b>, the id as a db_xref, and an evidence code as <b>post-text</b>.<br><b>properties</b> {<br>{<br>type comment,<br><b>heading</b> "GeneOntology",<br>source {<br>{<br>pre-text "Provided by",<br>anchor "RGD",<br>url "http://rgd.mcw.edu/"<br>}<br>},<br>comment {<br>{<br>type comment,<br><b>label</b> "Function",<br>comment {<br>{<br>type comment,<br><b>source</b> {<br>{<br>src {<br>db "GO",<br>tag id 3677<br>},<br><b>anchor</b> "DNA binding",<br><b>post-text</b> "evidence: IEA"<br>}<br>}<br>}<br>}<br>},</td></tr>
</table>